Framework para la Gestión de la Calidad de Datos en Gobierno Digital

1.0 - Diciembre, 2019

Este trabajo fue elaborado en colaboración con el Instituto de Computación y el Instituto de Agrimensura de la Facultad de Ingeniería - Universidad de la República, en el marco de un acuerdo entre AGESIC y la Fundación Julio Ricaldoni de la Facultad de Ingeniería para el Desarrollo de un Modelo de Referencia para Calidad de Datos en Gobierno Digital.

Tabla de Contenido

1 Introducción al Framework

Este capítulo presenta una introducción al Framework para la Gestión de Calidad de Datos en Gobierno Digital descripto en este trabajo. La Sección 1.1 brinda la motivación para el desarrollo del *framework*. La Sección 1.2 presenta los objetivos del *framework* y la Sección 1.3 introduce sus principales componentes. La Sección 1.4 describe el escenario de trabajo para el cual es aplicable el *framework* y la Sección 1.5 presenta problemas de calidad de datos comunes que se pueden dar en dicho escenario. Por último, la Sección 1.6 presenta la organización del resto del documento.

1.1. Motivación

Los avances en las tecnologías de la información y comunicación han llevado a que las organizaciones vinculadas a gobierno digital gestionen cantidades de datos cada vez más grandes. Esta realidad genera oportunidades interesantes en cuanto a su aprovechamiento, tanto para la operativa diaria de las organizaciones de gobierno así como para la toma de decisiones y planificación estratégica.

Sin embargo, estas oportunidades pueden verse limitadas por problemas de calidad de datos característicos de estos escenarios [BS16]. Por ejemplo, los datos de un mismo ciudadano se encuentran comúnmente en distintas bases de datos gestionadas de forma autónoma por diferentes organizaciones. Esto puede ocasionar que dos organizaciones o más manejen datos contradictorios de los ciudadanos. Asimismo, estos datos pueden haberse ingresado a lo largo de muchos años, utilizando procesos legados que pueden incluir varios pasos de ingreso de datos manual. Esto puede ocasionar que los datos no estén actualizados, tengan problemas de calidad ocasionados por su ingreso manual en los sistemas, y utilicen distintos formatos en las distintas organizaciones.

En este contexto, la calidad de datos en gobierno digital resulta de suma importancia, dado que tiene un fuerte impacto en la calidad de los servicios públicos que se brindan a los ciudadanos, así como en la definición de estrategias y políticas públicas.

Gestionar la calidad de datos en una organización implica distintas tareas, como medir, analizar, mejorar y controlar los distintos aspectos de la calidad de los datos. Resulta entonces de interés contar con mecanismos, guías y recomendaciones que asistan a las organizaciones vinculadas a gobierno digital en esta gestión.

En particular, la gestión de la calidad de datos en escenarios de gobierno digital implica varios desafíos, a causa de algunas de las características de estos escenarios:

- las organizaciones de gobierno manejan una gran cantidad de datos, potencialmente de todos los habitantes de un país [Boy11]
- la estructura de las bases de datos suele evolucionar de acuerdo a cambios en las normativas vigentes [Boy11]
- muchos de los datos que se almacenan tienen carácter probatorio y pueden servir de evidencia judicial, por lo que no pueden modificarse [Boy11]
- en general se apunta a garantizar que los ciudadanos suministren la misma información solo una vez a la administración pública [Tep17]
- las organizaciones tienen que a menudo utilizar datos gestionados por otras, por lo que no se tiene control directo sobre la calidad de estos datos [Tep17]

1.2. Objetivos del Framework

El objetivo general del *framework* es contribuir a la sistematización de la gestión de la calidad de datos en organizaciones vinculadas a gobierno digital en Uruguay, con el fin de mejorar la calidad de los datos que se generan y/o utilizan en el Estado Uruguayo. En este sentido, se apunta a que mediante la aplicación del *framework* se pueda contar con evidencia de mayor calidad para la toma de decisiones estratégicas, que permitan brindar mejores servicios y productos al ciudadano de forma más eficiente.

Para cumplir con este objetivo general, los objetivos específicos del *framework* son:

- proveer un marco conceptual que facilite comprender y manejar los distintos elementos involucrados en la gestión de calidad de datos en gobierno digital
- brindar un modelo de calidad de datos de referencia que facilite a las organizaciones definir modelos de calidad de datos específicos para sus escenarios de trabajo
- proponer un proceso que guíe a las organizaciones vinculadas a gobierno digital en cuanto a las actividades y roles relevantes para gestionar la calidad de los datos
- describir fundamentos teóricos relevantes para gestionar la calidad de datos en gobierno digital
- sugerir recursos de soporte (p. ej. bibliografía, herramientas, estándares) que sirvan de referencia y apoyo en la gestión de la calidad de datos en las organizaciones
- mostrar la utilización de los elementos anteriores en casos de estudio relevantes para gobierno digital y, en particular, para Uruguay

1.3. Componentes del Framework

El *framework* para la gestión de la calidad de datos está compuesto por un conjunto de elementos que se presentan gráficamente en la Figura 1.1.

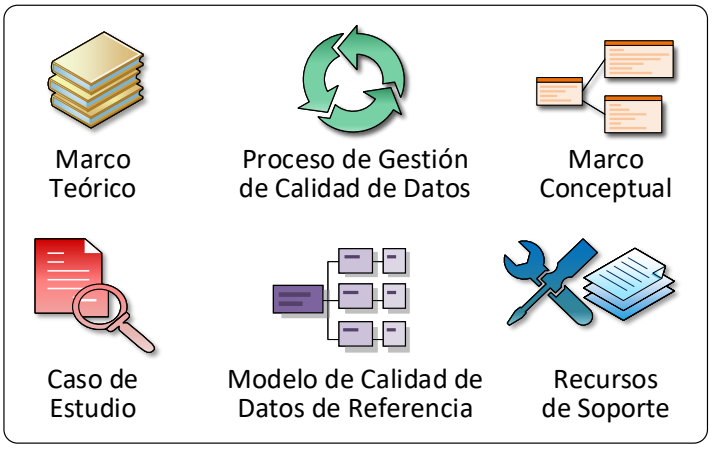


Figura 1.1: Componentes del Framework para la Gestión de la Calidad de Datos

El **Marco Teórico** proporciona fundamentos teóricos relevantes para gestionar la calidad de datos en gobierno digital así como para trabajar con el *framework*. En particular, el marco aborda temas de calidad de datos, modelos de calidad de datos, gestión de la calidad de datos, calidad de datos en gobierno digital y calidad en datos geográficos. El marco teórico se presenta en el Capítulo 2.

El **Marco Conceptual** define los principales conceptos relativos a la gestión de la calidad de datos en gobierno digital. Estos conceptos apuntan a facilitar la comprensión de la gestión de la calidad de datos en dichos contextos, por parte de los distintos actores involucrados. Asimismo, estos conceptos constituyen la base conceptual sobre la cual se apoya el *framework*. El marco conceptual se encuentra distribuido a lo largo de la mayoría de los capítulos entre el Capítulo 3 y el Capítulo 13.

El **Modelo de Calidad de Datos de Referencia** provee un conjunto extensible e instanciable de elementos de calidad de datos (p. ej. dimensiones, factores, métricas) con el fin de guiar y facilitar la definición de modelos de calidad de datos para escenarios de trabajo específicos. El modelo incluye cinco dimensiones de calidad de datos (i.e. Exactitud, Consistencia, Completitud, Unicidad y Frescura), dieciséis factores y más de cuarenta métricas. En particular, las métricas se pueden instanciar (i.e. configurar) para ser utilizadas en escenarios específicos. El modelo de referencia se presenta en el Capítulo 3 y se describe con más detalle en el Capítulo 16.

El **Proceso para la Gestión de Calidad de Datos** define los roles involucrados y las etapas a seguir para gestionar la calidad de datos en un escenario de gobierno digital. Dentro de los roles involucrados se destaca el comité de calidad de datos, liderado por un responsable de calidad de datos, que está a cargo de llevar adelante el proceso en un escenario específico. Las actividades incluyen, entre otras, la caracterización del escenario sobre el cual se va a aplicar el *framework*, examinar los datos objetivo y definir un modelo de calidad de datos. El proceso para la gestión de la calidad de datos se presenta en el Capítulo 5. Además, entre el Capítulo 6 y el Capítulo 13 se detallan cada una de las etapas de este proceso.

Los **Recursos de Soporte** son un conjunto extensible de artefactos que brindan soporte o sirven de referencia para la gestión de la calidad de datos así como para la aplicación del *framework*. Estos recursos de soporte incluyen, entre otros, herramientas, técnicas, servicios y estándares. Los recursos de soporte se presentan en el Capítulo 14.

Por último, el **Caso de Estudio** plantea un escenario de trabajo concreto para motivar, ejemplificar y guiar la gestión de la calidad de datos en estos contextos, en particular, cuando se utiliza el *framework*. El caso plantea una realidad que involucra, varias organizaciones de gobierno, distintos tipos de sistemas y aplicaciones, así como distintos tipos de datos (p. ej. alfanuméricos, geográficos). El caso de estudio se presenta en el Capítulo 4. La aplicación del *framework* en dicho caso de estudio se describe parcialmente entre el Capítulo 6 y el Capítulo 13, así como de forma unificada en el Capítulo 15.

1.4. Escenario de Trabajo

El escenario de trabajo en el cual es aplicable el *framework* se enmarca en un contexto de gobierno digital en el cual distintas organizaciones (p. ej. organismos públicos) llevan a cabo procesos de negocio y colaboran entre sí con el fin de ofrecer servicios públicos a los ciudadanos. La Figura 1.2 presenta, a modo ilustrativo, algunos de los principales elementos que pueden encontrarse en un escenario de trabajo tipo.

Las organizaciones se apoyan en aplicaciones y sistemas para llevar a cabo sus procesos de negocio, que pueden ser tanto implícitos como estar explícitamente implementados en sistemas especializados. Estas aplicaciones y sistemas utilizan colecciones de datos de distintos tipos (p. ej. bases de datos relacionales, archivos de texto) que almacenan la información de las entidades de negocio relevantes para la organización (p. ej. certificados, empresas, ciudadanos, beneficios). Las entidades de negocio (p. ej. ciudadano) tienen atributos (p. ej. nombre, dirección) que pueden ser de distinto tipo (p. ej. alfanuméricos, geográficos, imágenes).

Las organizaciones ofrecen un portal a través del cual interactúan con los ciudadanos, permitiéndoles participar en procesos de negocio (p. ej. realizar trámites) así como acceder a los datos que almacenan. Los organizaciones también interactúan entre sí (p. ej. para consultar datos) a través de servicios de negocio.

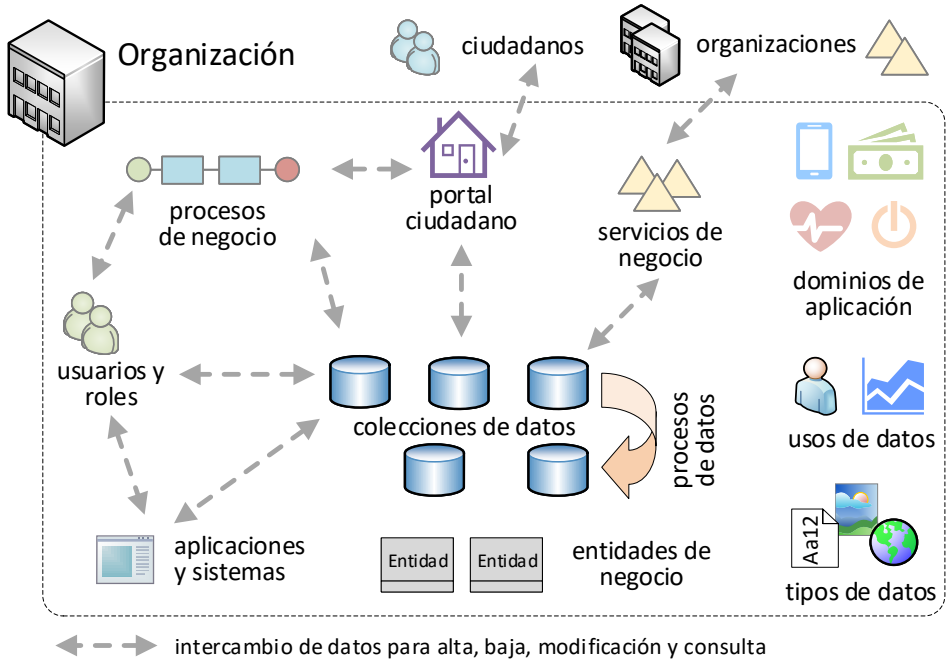


Figura 1.2: Principales Elementos del Escenario de Trabajo

Los datos en las colecciones son creados, borrados, modificados y consultados de forma manual (i.e. con intervención de una persona) o automática por distintos tipos de clientes de datos. Los clientes de datos pueden ser aplicaciones, sistemas, procesos de negocio, servicios de negocio, portal ciudadano, usuarios con distintos roles (p. ej. técnico, gerente) y procesos de datos (p. ej. procesos ETL[1] para cargar un *data warehouse*).

Los organizaciones pueden operar en distintos dominios de aplicación (p. ej. salud, energía, telecomunicaciones, finanzas) y pueden utilizar los datos con distintos fines (p. ej. operativa de la organización, toma de decisiones).

[1] extracción, transformación y carga

1.5. Problemas Comunes de Calidad de Datos

Esta sección presenta ejemplos comunes de problemas vinculados a la calidad de datos que se pueden dar en un escenario de gobierno digital, como el que se presenta en la Sección 1.4.

1.5.1. Calidad de Datos en el Marco de una Aplicación

Existen varios problemas de calidad de datos que pueden surgir en el marco de una única aplicación que almacena datos en una única colección de datos (p. ej. una base de datos relacional). La Figura 1.3 presenta algunos de estos problemas, los cuales hacen referencia al formato, unicidad y exactitud de los datos.

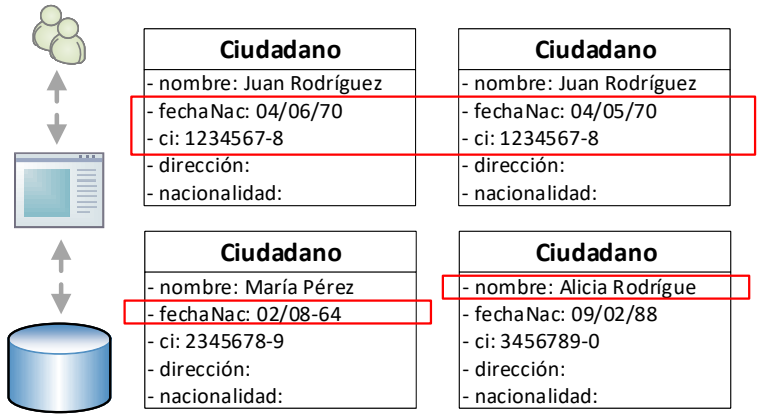


Figura 1.3: Calidad de Datos en una Aplicación

Con respecto al formato de los datos se puede observar que la fecha de nacimiento de la ciudadana María Pérez no sigue el formato de fecha establecido (i.e. dd/mm/aa).

En relación a la unicidad de los datos se puede observar que los datos del ciudadano Juan Rodríguez se encuentran duplicados en la colección de datos. Además, estos datos duplicados presentan contradicciones entre sí en cuanto a la fecha de nacimiento del ciudadano.

Por último, con respecto a la exactitud de los datos se puede observar que los datos del nombre de la ciudadana llamada Alicia Rodríguez no son sintácticamente correctos.

1.5.2. Calidad de Datos en un Conjunto de Aplicaciones

Los problemas de calidad en los datos tienden a crecer rápidamente cuando se consideran varias aplicaciones que pueden utilizar una o más colecciones de datos. En la Figura 1.4 se observa cómo dos aplicaciones, que utilizan su propia colección de datos, manejan datos de un mismo ciudadano que son incompletos y contradictorios entre sí.

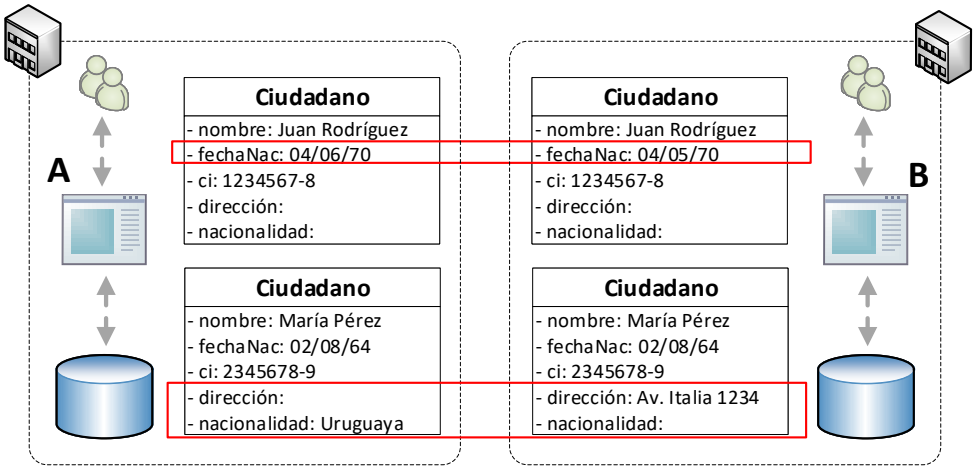


Figura 1.4: Calidad de Datos en un Conjunto de Aplicaciones

En particular, la fecha de nacimiento del ciudadano Juan Rodríguez que maneja la aplicación A es distinta a la que maneja la aplicación B. Asimismo, la aplicación A no tiene datos de la dirección de la ciudadana María Pérez mientras que la aplicación B sí. De forma similar, la aplicación B no tiene datos de la nacionalidad de la ciudadana María Pérez mientras que la aplicación A sí.

1.5.3. Calidad en Intercambio de Datos

Aún en el caso ideal en que las colecciones de datos que manejan las aplicaciones no presenten problemas de calidad, estos problemas pueden surgir al intercambiar datos entre distintas aplicaciones de la misma o de diferentes organizaciones. La Figura 1.5 presenta un ejemplo de esta situación.

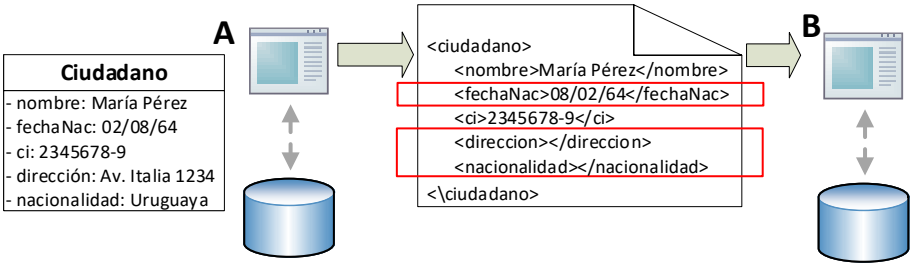


Figura 1.5: Calidad en Intercambio de Datos

En este caso, si bien los datos de la ciudadana María Pérez no tienen problemas de calidad en la colección de datos de la aplicación A, al intercambiarlos con la aplicación B se originan los siguientes problemas:

- se intercambian los valores del día y mes en la fecha de nacimiento
- los datos de la dirección y nacionalidad no se incluyen en los datos intercambiados

Como consecuencia, la aplicación B recibe datos con problemas de calidad que fueron originados por los procesos de intercambio de datos.

1.6. Organización del Documento

El resto del documento se organiza de la siguiente forma:

- El Capítulo 2 presenta marco teórico relevante para gestionar la calidad de datos en gobierno digital así como para la utilización del *framework*.
- El Capítulo 3 describe el modelo de calidad de datos de referencia.
- El Capítulo 4 describe el caso de estudio utilizado para mostrar la aplicación del *framework*.
- El Capítulo 5 presenta el proceso para la gestión de la calidad de datos, el cual consiste de siete etapas.
- El Capítulo 6 detalla parte de la primera etapa del proceso, que corresponde a la caracterización técnica y de negocio del escenario en el cual se va a aplicar el *framework*.
- El Capítulo 7 detalla otra parte de la primera etapa del proceso, que propone la caracterización de calidad de datos del escenario en el cual se va a aplicar el *framework*.
- El Capítulo 8 detalla la segunda etapa del proceso, en la que se busca examinar los datos objetivo.
- El Capítulo 9 detalla la tercera etapa del proceso, que corresponde a la definición de la estrategia de gestión de calidad a utilizar.
- El Capítulo 10 detalla la cuarta etapa del proceso, que se enfoca en la definición del modelo de calidad de datos para el escenario.
- El Capítulo 11 detalla la quinta etapa del proceso, que consiste en medir y evaluar la calidad de datos en base al modelo de calidad definido.
- El Capítulo 12 detalla la sexta etapa del proceso, que apunta a determinar las causas de los problemas de calidad encontrados.
- El Capítulo 13 detalla la séptima etapa del proceso, que consiste en definir, ejecutar y evaluar un plan de mejora para la calidad de los datos en el escenario.
- El Capítulo 14 presenta los recursos de soporte sugeridos.
- El Capítulo 15 describe de forma unificada la aplicación del *framework* en el caso de estudio.
- El Capítulo 16 presenta con más detalle las dimensiones del modelo de calidad de datos de referencia descripto en el Capítulo 3.

2 Marco Teórico

En este capítulo se presenta marco teórico relevante para gestionar la calidad de datos en gobierno digital así como para la utilización del *framework*.

En la Sección 2.1 se presenta el concepto de calidad de datos. La Sección 2.2 hace foco en modelos de calidad y la Sección 2.3 brinda detalles de la tarea de gestionar la calidad de datos. La Sección 2.4 presenta las particularidades de la calidad de datos en un contexto de gobierno digital. Por último, la Sección 2.5 presenta aspectos de la calidad en datos geográficos.

2.1. Calidad de Datos

Calidad de Datos es un área de investigación muy amplia, que implica muchos aspectos, problemas y desafíos. Tiene además una enorme relevancia para la industria, debido a su gran impacto en los sistemas de información en todos los dominios de aplicación. El término «calidad de datos» se utiliza con referencia a un conjunto de características que deben poseer los datos, tales como su correctitud y su grado de actualización, entre otros [SC02]. Las consecuencias de la mala calidad de los datos a menudo se experimentan en la vida cotidiana (p. ej. una dirección incorrecta o duplicada).

Calidad de datos también se define como la capacidad de cumplir con los requerimientos definidos para el uso de los datos. De esta forma, los datos carecen de calidad en la medida en que no satisfacen los requerimientos [Ols03] y, por este motivo, la calidad de datos depende tanto del uso que se le vaya a dar a los datos, como de los datos en sí. Otra definición similar ampliamente adoptada es la que considera que la calidad de datos es la adecuación de los datos para su uso (i.e. *fitness for use*), independientemente de si los requerimientos están explícitamente definidos [WS96] [SC02] [Nee05] [Str97] [Cha05] [TB98].

La calidad de datos se asocia a un conjunto de dimensiones que agrupan propiedades o características de calidad [SC02]. Las dimensiones más frecuentemente consideradas para calidad de datos son exactitud, actualidad, completitud y consistencia [Fox94]. Si bien no hay acuerdo sobre todo el conjunto de dimensiones a considerar ni sobre sus significados [SC02], existen algunos consensos en lo que refiere a estas dimensiones.

2.2. Modelos de Calidad de Datos

Como se mencionó previamente, la calidad de datos se puede caracterizar de acuerdo a dimensiones que ayudan a calificar los datos [Etc08] [Ako07]. La calidad de datos es entonces un concepto multi-facético, dado que se representa por un conjunto de dimensiones que abordan diferentes aspectos de los datos. Como se presenta en la Figura 2.1, siguiendo este enfoque es posible definir una jerarquía de conceptos de calidad.

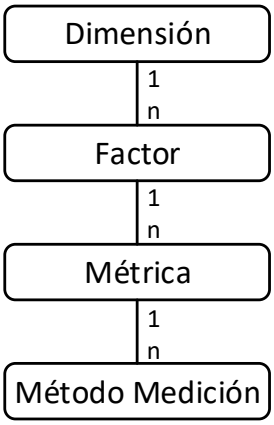


Figura 2.1: Jerarquía de Conceptos de Calidad de Datos

En este enfoque cada dimensión captura una faceta o característica de la calidad y puede verse como un conjunto de factores que tienen un mismo propósito. Un factor de calidad representa un aspecto particular de una dimensión y puede ser medido con distintas métricas. Una métrica es un instrumento que define la forma de medir un factor de calidad y puede ser medida por diferentes métodos. Un método es un proceso que implementa una métrica y permite obtener medidas de calidad para la misma.

Para abordar la calidad de datos en un sistema de información es necesario definir un modelo de calidad adecuado a las necesidades y prioridades de los consumidores de los datos. El modelo de calidad de datos define qué dimensiones de calidad se consideran, sobre qué datos se aplican dichas dimensiones y cómo se miden. Este modelo es el que guía la gestión de la calidad para un conjunto de datos específico.

A modo de ejemplo, la Tabla 2.1 presenta un posible modelo de calidad definido para los datos de ventas, productos y clientes almacenados en una base de datos relacional.

Tabla 2.1: Ejemplo de Modelo de Calidad de Datos

Dimensión	Factor	Métrica genérica	Métrica instanciada sobre:
Exactitud	Correctitud sintáctica	M1: Formato gran: celda tipo-res: {0,1}	Clientes.ci Clientes.sexo Productos.prov Ventas.importe
	Precisión	M2: CantDecim gran: columna tipo-res: {0,1}	Productos.cant-stock Ventas.importe
Completitud	Cobertura	M3: RatioCobertura gran: tabla tipo-res: [0,1]	Clientes

En particular, el modelo incluye dos dimensiones de calidad: Exactitud y Completitud. Para la dimensión Exactitud interesa considerar dos factores de calidad: Correctitud Sintáctica y Precisión. Por otro lado, para la dimensión Completitud solo interesa considerar el factor de calidad Cobertura.

Para el factor de calidad Correctitud Sintáctica, se define una métrica genérica M1 denominada *Formato*. Esta métrica aplica a una celda de una tabla de la base de datos (i.e. tiene granularidad celda). El resultado de M1 es un valor 0 o 1.

Por otro lado, para el factor de calidad Precisión se define la métrica genérica M2 denominada *CantDecim*. Esta métrica aplica a una columna de la base de datos (i.e. tiene granularidad columna). Al igual que M1, el resultados de M2 es un valor 0 o 1.

En el caso del factor de calidad Cobertura, se define la métrica M3 denominada *RatioCobertura*. Esta métrica aplica a una tabla de la base de datos (i.e. tiene granularidad tabla). El resultado de M3 es un valor entre 0 y 1.

Estas métricas genéricas se utilizan para definir métricas instanciadas, las cuales determinan sobre qué elementos de la base de datos (p. ej. celdas, columnas, tablas), se van a aplicar las métricas genéricas. Por ejemplo, la métrica genérica M1 se instancia sobre las celdas correspondientes a los atributos ci y sexo de la tabla Clientes, así como sobre las celdas correspondientes a los atributos prov e importe de las tablas Productos y Ventas, respectivamente.

2.3. Gestión de Calidad de Datos

La gestión de la calidad de datos es la tarea de medir, analizar, mejorar y controlar, entre otras, los distintos aspectos de calidad de los datos que son de interés para un escenario específico. Esta tarea implica un conjunto de metodologías, técnicas y herramientas para manejar la calidad de los datos, en una organización o en varias que están cooperando.

2.3.1. Elementos para la Gestión de la Calidad de Datos

La Figura 2.2 presenta los principales elementos que dan soporte a la gestión de la calidad de datos.

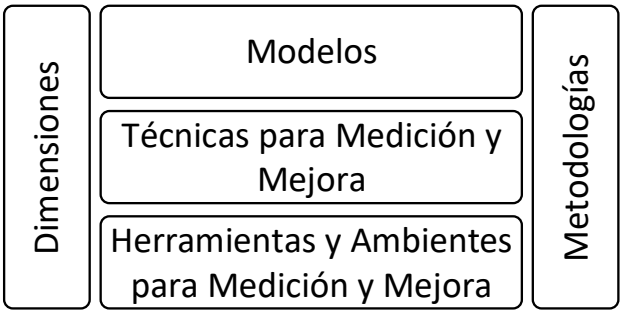


Figura 2.2: Elementos para la Gestión de la Calidad de Datos [BS16]

En primer lugar, es necesario contar con un modelo de calidad en el cual se definen las dimensiones adecuadas para el conjunto de datos de interés. Para construir dicho modelo, se aplican metodologías que apoyan, en primer lugar, en la identificación y clasificación de los problemas de calidad de datos. A partir de los tipos de problemas de calidad que se identifican, se seleccionan técnicas, y herramientas adecuadas para dichas técnicas, para la medición y la mejora de la calidad de los datos.

2.3.2. Etapas de la Gestión de la Calidad de Datos

La Figura 2.3 presenta distintas etapas en la gestión de la calidad de datos.

En primer lugar, se observa una etapa de análisis de procesos de negocio involucrados que permite identificar el conjunto de datos relevantes para el escenario. Seguidamente, la etapa *Data Profiling* permite conocer estos datos para los cuales se va a construir el modelo de calidad. Esta etapa de reconocimiento de los datos, permite una primera estimación de la calidad de los mismos. En base a los resultados obtenidos de esta etapa, así como de un primer análisis de las causas de la mala calidad, se define el modelo de calidad.

El modelo de calidad permite analizar las causas de mala calidad, por lo que la construcción del modelo y el análisis de las causas de la mala calidad, es un proceso paralelo e iterativo.

El modelo de calidad permite realizar la medición y evaluación de la calidad de los datos. En base a los resultados de la medición y evaluación, se pueden determinar acciones para la mejora de la calidad (p. ej. re-estructuración del sistema, limpieza de los datos).

Finalmente y para mantener un nivel de calidad de datos de acuerdo al requerido por la o las organizaciones, la gestión requiere del monitoreo constante de la calidad. El monitoreo de la calidad puede requerir ajustes al modelo de calidad así como nuevas mediciones y evaluaciones de la calidad de los datos.

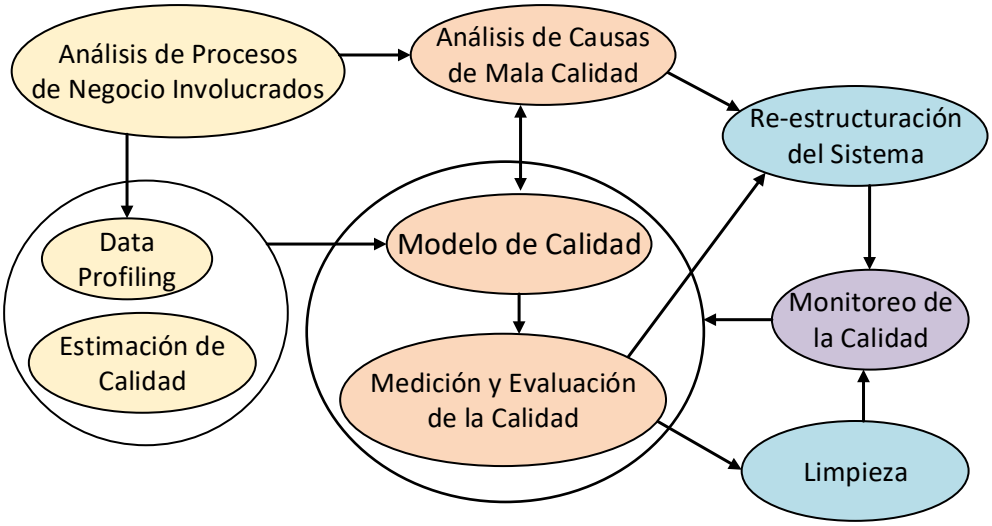


Figura 2.3: Etapas de la Gestión de la Calidad de Datos

2.3.3. Metodologías para la Gestión de la Calidad de Datos

Como se presentó en la Figura 2.2, las Metodologías son elementos importantes para la gestión de la calidad de datos.

Las metodologías de calidad de datos son un conjunto de pautas y técnicas que, a partir de un conjunto de datos relacionados con una realidad o dominio de interés, definen un proceso para evaluar y mejorar la calidad de los datos a través de diferentes etapas [BS16].

Las metodologías se pueden clasificar en las siguientes categorías [BS16]:

- **Completas:** presentan etapas de evaluación y mejora; además abordan cuestiones técnicas y económicas.
- **De auditoría:** se centran en la etapa de evaluación y brindan un apoyo limitado a la etapa de mejora.
- **Operativas:** se centran en los problemas técnicos de las etapas de evaluación y mejora, pero no abordan problemas económicos.
- **Económicas:** se centran en la evaluación de costos.

En particular, la metodología *Complete Data Quality Methodology* (CDQM) presenta un balance entre la completitud y la factibilidad práctica de los procesos de mejora [BS16]. Si bien esta metodología se enfoca principalmente en datos relacionales estructurados, también se extiende a otros tipos de datos semi-estructurados.

2.4. Calidad de Datos en Gobierno Digital

La calidad de datos en un contexto de gobierno digital resulta de suma importancia, dado que tiene un fuerte impacto en la calidad de los servicios que se brindan a los ciudadanos, así como en las políticas y estrategias que se definen.

En particular, la calidad de los datos que refieren a los ciudadanos resulta clave para proporcionarles buenos servicios. Por ejemplo, en los sistemas de información de salud, los datos disponibles, confiables, oportunos y válidos son un requisito para la prestación de servicios de salud de alta calidad. En un contexto de gobierno digital se debe apuntar además a que los ciudadanos suministren una misma información solo una vez a la administración pública.

Experiencias en algunos países indican que para diseñar un modelo de calidad de datos para organizaciones vinculadas a gobierno digital deben tenerse en cuenta sus particularidades [Tep17][Boy11]. Concretamente, el modelo de calidad de datos, así como las características de calidad que incluya, deben cumplir con los siguientes requisitos [Tep17]:

- las características de calidad deben ser importantes para las organizaciones
- las características deben poder ser influenciadas por la organización que es propietaria de los datos
- las características deben cubrir (tanto como sea posible) todos los aspectos de calidad de datos relevantes, teniendo en cuenta además que el número de características debe mantenerse bajo para permitir un uso eficiente del modelo de calidad
- las características deben ser únicas (no deben duplicarse entre sí)
- las características deben ser medibles
- en el caso que haya más de un modelo de calidad de datos, deben tener prioridad aquellos desarrollados por organismos referentes en el área de gobierno digital

2.5. Calidad en Datos Geográficos

En esta sección se brinda una introducción a los datos geográficos y se describen problemas de calidad comunes en los mismos.

2.5.1. Datos Geográficos

Los datos geográficos son datos que pueden relacionarse con una ubicación en la tierra [Ber12]. Dicha ubicación puede representarse de acuerdo a modelos de datos vectoriales o de matriz (conocido como Modelo Ráster). Para representar elementos complejos en la superficie de la tierra en el modelo de datos vectorial se distinguen varios tipos de geometrías, siendo los básicos: puntos, líneas y polígonos.

A su vez para representar las ubicaciones en la tierra se utilizan diferentes sistemas de coordenadas. El más conocido es Latitud-Longitud donde las coordenadas se expresan como grados. También hay otros sistemas de coordenadas llamados planos donde se usan coordenadas X,Y. El sistema de coordenadas utilizado en Uruguay es el SIRGAS ROU 98 o WGS98 proyectados a UTM. Las coordenadas X,Y representan medidas en metros desde determinados marcos de referencia. Existen transformaciones matemáticas para convertir datos de un sistema de coordenadas a otro.

Los datos geográficos se pueden obtener a partir de diversos procesos y con diferentes instrumentos (p. ej. estación total, tecnologías basadas en láser LIDAR, sensores, GPS, fotos satelitales). Esto hace que en los procesos de generación de los datos se puedan introducir errores por variaciones en la precisión de los instrumentos y procesos utilizados.

2.5.2. Problemas de Calidad Comunes en Datos Geográficos

Al trabajar con datos geográficos se pueden presentar distintos problemas de calidad referentes a la exactitud, consistencia y duplicación.

Por ejemplo, pueden surgir incompatibilidades al trabajar con conjuntos de datos geográficos de distintas fuentes que no tienen los mismos niveles de calidad. La Figura 2.4 presenta un ejemplo de esta situación, en el cual se pueden observar dos conjuntos de datos que, por no tener los mismos niveles de exactitud posicional, no coinciden al superponerlos (interoperabilidad geográfica) como se presenta en la Figura 2.5.

Figura 2.4: Conjuntos de Datos Geográficos de Distintas Fuentes

Figura 2.5: Superposición de Conjuntos de Datos Geográficos de Distintas Fuentes

También pueden surgir problemas referentes a la consistencia de los datos en relación a reglas de dominio. Por ejemplo, los permisos de extracción de materiales para la construcción (p. ej. arena, canto rodado) solo pueden otorgarse sobre cursos de agua de dominio público. La Figura 2.6 muestra que hay dos permisos de construcción que no están ubicados sobre cursos de agua, por lo que rompen la regla antes mencionada.

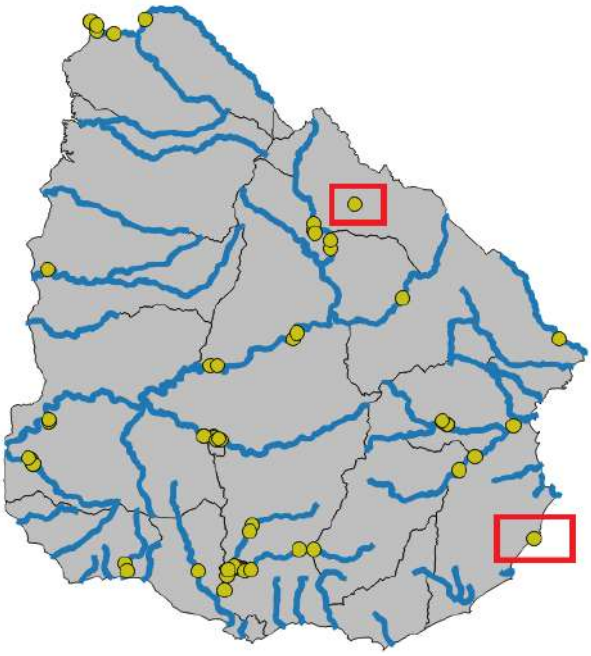


Figura 2.6: Consistencia con Reglas de Dominio

Por último, otro problema que presentan los datos espaciales es la omisión o comisión[1] de objetos en el conjunto de datos espaciales. Esto se debe principalmente a dos motivos: el paso del tiempo y los métodos de generación a través de fuentes secundarias. Por ejemplo, un conjunto de datos geográficos vectoriales puede no corresponderse completamente con lo reflejado en la imagen a partir del cual se generaron, debido a falta de datos geográficos o inclusión de datos geográficos que no existen en la realidad, entre otros.

La Figura 2.7 presenta un ejemplo de esta situación en donde a partir de una imagen satelital se generó un conjunto de datos geográficos vectoriales de las construcciones representadas como puntos. En este caso se pueden observar zonas donde existen construcciones que no fueron reconocidas (en rojo), así como construcciones que no existen en la realidad (en magenta).

[1] objetos geográficos presentes en los datos, pero no en la realidad [IDE18]

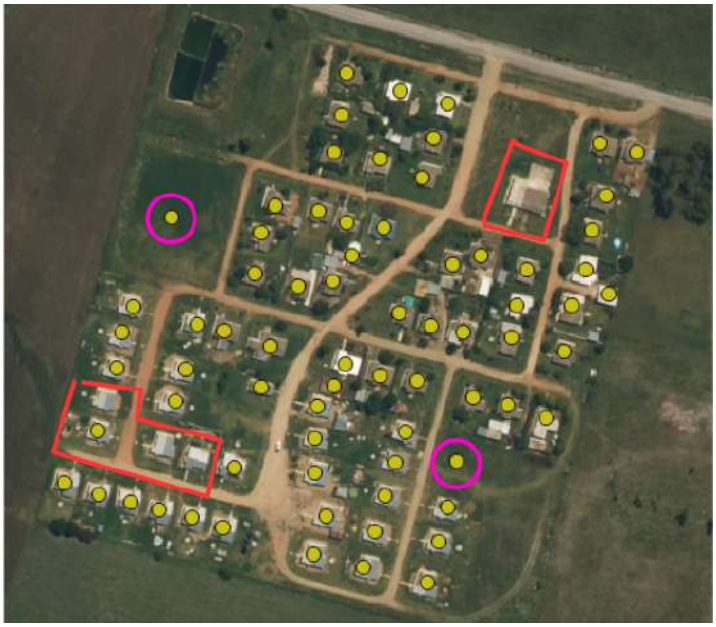


Figura 2.7: Discrepancias entre Datos Vectoriales e Imagen Satelital

3 Modelo de Calidad de Datos de Referencia

Como se mencionó en la Sección 2.2, para la gestión de la calidad de datos es clave contar con un modelo de calidad que considere los requerimientos, problemas y prioridades del escenario. Este capítulo propone un Modelo de Calidad de Datos de Referencia que apunta a guiar la definición de modelos de calidad de datos para escenarios específicos.

La Sección 3.1 describe el marco conceptual asociado a los elementos que componen los modelos de calidad de datos. La Sección 3.2 presenta una descripción general del modelo de referencia propuesto en base a estos elementos. Entre la Sección 3.3 y la Sección 3.7 se detallan los elementos del modelo de referencia, organizados por las dimensiones que incluye.

3.1. Marco Conceptual Asociado

Esta sección describe los principales conceptos asociados a los elementos que componen los modelos de calidad de datos. Estos conceptos son la base para la especificación del modelo de calidad de datos de referencia, así como de los modelos de calidad específicos que se definan para escenarios concretos.

3.1.1. Elementos del Modelo

La Figura 3.1 presenta conceptos relacionados a elementos de modelos de calidad. En particular, estos conceptos permiten la especificación de: modelos de calidad de datos y sus elementos (i. e. dimensiones, factores, métricas y métodos).

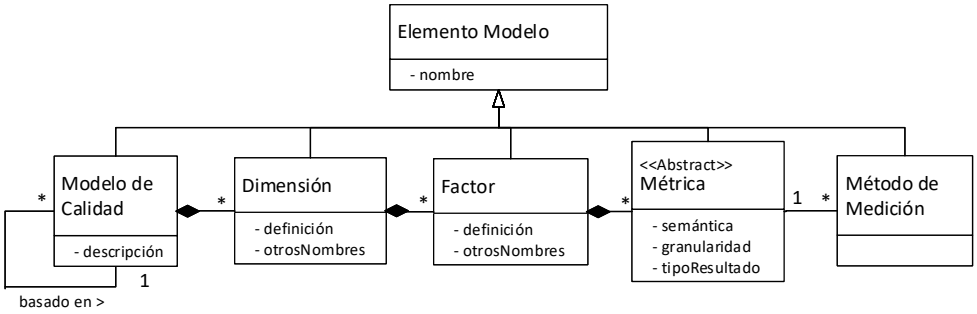


Figura 3.1: Elementos del Modelo

Un Modelo de Calidad está compuesto por un conjunto de dimensiones que representan las cuestiones de calidad de datos relevantes para un determinado escenario o conjunto de escenarios. Tomando como base lo presentado en la Sección 2.2:

- una Dimensión captura una faceta a alto nivel de la calidad de datos (p. ej. exactitud)
- un Factor representa un aspecto particular de una dimensión de calidad (p. ej. correctitud sintáctica)
- una Métrica es un instrumento que define la forma de medir un factor de calidad (p. ej. porcentaje de datos sintácticamente correctos)
- un Método de Medición es un proceso que implementa una métrica y se encarga de tomar medidas para la misma

Para definir una métrica es necesario especificar su semántica (i.e. qué se mide), granularidad (i.e. sobre qué se mide) y el tipo de resultado (p. ej. porcentaje). La Tabla 3.1 presenta los tipos de granularidad que considera el *framework* y su correspondencia con el modelo relacional.

Tabla 3.1: Tipos de Granularidad

Granularidad	**En Modelo Relacional**
instanciaAtributo	celda
atributo	columna
conjuntoAtributos	conjunto de columnas
instanciaEntidad	tupla
entidad	tabla
conjuntoEntidades	conjunto de tablas
colección	base de datos
conjuntoColecciones	conjunto de bases de datos
organización	organización
conjuntoOrganizaciones	conjunto de organizaciones

El Ejemplo 1 presenta la definición de una métrica denominada RatioNoNulos-DireccionCliente que mide el porcentaje de clientes con valores no nulos en la dirección.

Ejemplo 1
Nombre: RatioNoNulos-DireccionCliente
Semántica: Porcentaje de clientes que tienen un valor no nulo para la dirección.
Granularidad: atributo
Tipo de Resultado: Intervalo real [0.0, 1.0]

3.1.2. Métricas Genéricas y Específicas

Como se presenta en la Figura 3.2, el *framework* distingue entre Métricas Genéricas y Métricas Específicas.

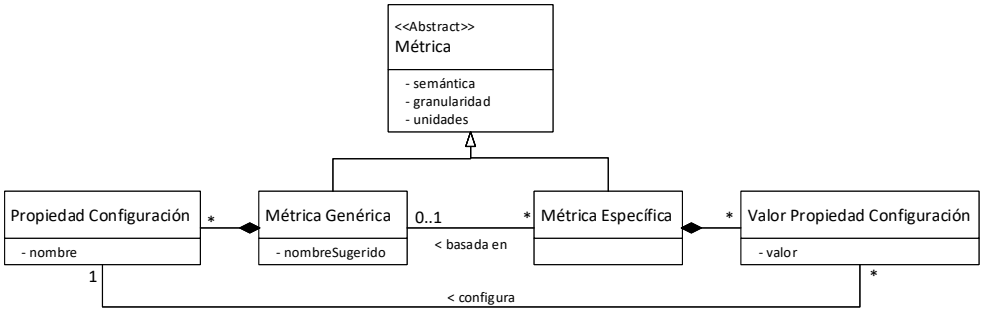


Figura 3.2: Métricas Genéricas y Específicas

Una Métrica Genérica representa una familia de métricas (p. ej. RatioNoNulos) con características similares. El objetivo de estas métricas es guiar, agilizar y uniformizar la creación de nuevas métricas, brindando definiciones generales que pueden ser refinadas por métricas más específicas (p. ej. RatioNoNulos-Direccion, RatioNoNulos-Apellido). El refinamiento se realiza estableciendo valores para las Propiedades de Configuración (p. ej. atributo) que define la métrica genérica, así como especificando una semántica más concreta.

El Ejemplo 2 presenta la definición de una métrica genérica que indica si el valor de un atributo cumple con un formato determinado y define dos propiedades de configuración: atributo y estándar o diccionario.

Ejemplo 2
Nombre: Formato
Semántica: Indica si el valor de un atributo cumple con el formato definido para ese atributo según algún estándar o diccionario
Granularidad: instanciaAtributo
Tipo de Resultado: Boolean
Nombre Sugerido: Formato(atributo, estándar o diccionario)
Propiedades de Configuración: atributo, estándar o diccionario

A partir de esta métrica genérica es posible definir varias Métricas Específicas, por ejemplo:

- Formato(país, ISOAlpha2)
- Formato(país, ISOAlpha3)
- Formato(cédula, DNIC)

En particular, el Ejemplo 3 presenta la definición de la métrica específica «Formato(país, ISOAlpha3)», la cual está basada en la métrica genérica del Ejemplo 2 e indica si el valor del atributo país está en el formato ISO 3166-1 alpha-3 (3 letras).

Ejemplo 3
Nombre: Formato(Pais, ISOAlpha3)
Semántica: Indica si el código de un país está en el formato ISO 3166-1 alpha-3.
Granularidad: instanciaAtributo
Tipo de Resultado: Boolean
Propiedades de Configuración: atributo = país, estándar = ISO 3166-1 alpha-3

Las métricas genéricas y específicas tienen como objetivo agilizar la definición de un modelo de calidad de datos para un escenario particular, dado que son métricas predefinidas aplicables a distintos escenarios y que abordan temas de calidad comunes. A modo de ejemplo, la métrica «Formato(País,ISOAlpha3)» puede ser utilizada en varios escenarios de trabajo donde se tenga el requerimiento de que los países tienen que especificarse con el formato ISO 3166-1 alpha-3.

3.1.3. Aplicabilidad de Elementos del Modelo

Como se presenta en la Figura 3.3, los elementos del modelo (p. ej. dimensiones, factores, métricas) pueden aplicar a uno o más tipos de datos (p. ej. alfanumérico, imagen, geográfico), tipos de colecciones de datos (p. ej. base de datos relacional, base de datos documental), dominios de aplicación (p. ej. salud, transporte, energía) y usos de datos (p. ej. operativa, toma de decisiones).

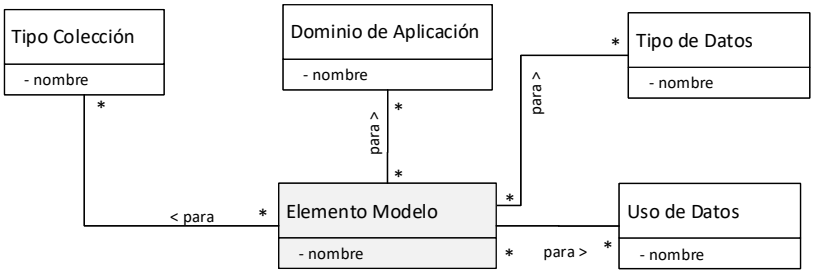


Figura 3.3: Aplicabilidad de Elementos del Modelo

Por ejemplo, un modelo puede tener un factor Consistencia-Topológica que aplica únicamente al tipo de datos Geográfico. Asimismo, un modelo puede tener una métrica Consistencia-Enfermedad-Sexo que aplica únicamente al dominio de la Salud.

3.1.4. Definición de Métricas por Agregación

Las métricas definidas por agregación, son métricas que pueden calcularse a partir de métricas ya existentes. En particular, dadas dos métricas M1 y M2 para el mismo factor, si la semántica de M2 puede expresarse como un cálculo que agrega un conjunto de resultados de M1, entonces M2 es una métrica agregada de M1. Si M1 está definida para la mínima granularidad posible, entonces M1 es una métrica atómica.

Por ejemplo, si el modelo de calidad de referencia especifica la métrica NoNulo, cuya granularidad es instanciaAtributo y cuyo resultado es «true» si la instancia del atributo tiene un valor no nulo, es posible definir una nueva métrica agregada, denominada RatioNoNulos. Esta nueva métrica tiene granularidad atributo y calcula la proporción de valores no nulos según la métrica NoNulo, sobre el total de instancias de atributo sobre la que se realizó la medición.

En la Tabla 3.2 se muestran los casos más comunes de cambios en la granularidad entre una métrica atómica y una métrica agregada.

Tabla 3.2: Agregación de Métricas según Granularidad

Granularidad Métrica Atómica	Granularidad Métrica Agregada
instanciaAtributo	atributo
instanciaEntidad	entidad
atributo	entidad
entidad	conjuntoEntidades
entidad	colección
colección	conjuntoColecciones
colección	organización
organización	conjuntoOrganizaciones

En la Tabla 3.3 se presentan cuatro tipos muy utilizados de métricas agregadas.

Tabla 3.3: Tipos de Métricas Agregadas

Métrica Agregada	**Semántica**
Ratio	Proporción de valores en «true» sobre el total de valores medidos.
RatioUmbral	Proporción de valores iguales o mayores a un umbral, sobre el total de valores medidos. El umbral pertenece al intervalo real [0, 1].
Promedio	Valor promedio de todos los valores medidos.
PromedioPonderado	Promedio ponderado de los valores medidos, en donde cada valor se multiplica por un coeficiente o peso.

En la Tabla 3.4 se muestra, por ejemplo, cómo a partir de la definición de una métrica atómica (NoNulo) que se calcula a nivel de instancia de atributo, puede definirse una nueva métrica agregada de tipo Ratio (RatioNoNulos) que se calcula a nivel de atributo.

Tabla 3.4: Definición de una Métrica Agregada a partir de una Métrica Atómica

Métrica Atómica	**Métrica Agregada**
Nombre: **NoNulo** ***Semántica:*** Indica si una instancia de atributo tiene un valor no nulo. Puede ser necesario especificar un diccionario con todos los valores del atributo que se consideran nulos o vacíos, cuando existe más de una posibilidad. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean ***NombreSugerido:*** NoNulo(Atributo) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. DiccionarioValorNoNulos(Atributo)	***Nombre:*** **RatioNoNulos** ***Semántica:*** Proporción de valores no nulos según la métrica NoNulo, sobre el total de instancias de entidad sobre las que se realizó la medición. ***Granularidad:*** atributo. ***TipoResultado:*** Intervalo real [0, 1]. ***NombreSugerido:*** RatioNoNulos(Atributo) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. DiccionarioValoresNulos(Atributo)

Con la finalidad de simplificar el modelo de referencia, solo se incluyen algunas métricas agregadas de uso frecuente, pero es posible definir otras métricas agregadas a partir de métricas atómicas, siguiendo un procedimiento análogo al explicado anteriormente.

3.2. Descripción General del Modelo de Referencia

El Modelo de Calidad de Datos de Referencia se especifica en base a los conceptos de modelos de calidad de datos que se describen en la Sección 3.1 (p. ej. dimensiones, factores, métricas) y que pueden relacionarse de forma jerárquica en varios niveles.

En particular, el modelo de referencia propuesto tiene en un primer nivel cinco dimensiones: Exactitud, Consistencia, Completitud, Unicidad y Frescura. El segundo y tercer nivel están formados por los factores de cada dimensión y por las métricas de cada factor, respectivamente. En la Figura 3.4 se resumen las dimensiones y los factores del modelo propuesto, los cuales se describen con más detalle entre la Sección 3.3 y la Sección 3.7.

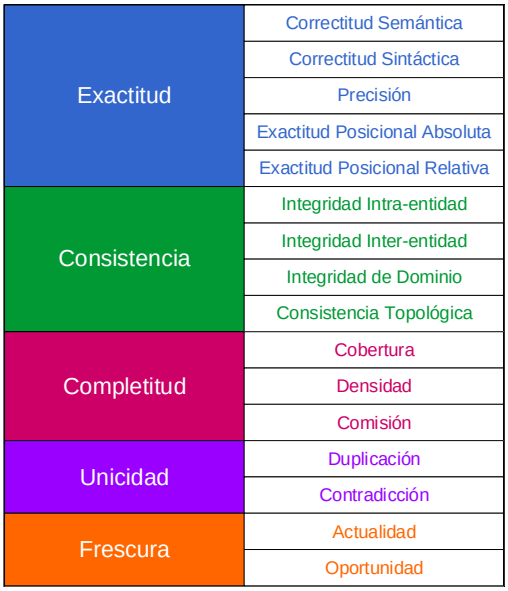


Figura 3.4: Dimensiones y Factores del Modelo de Calidad de Datos de Referencia

Como se mencionó en la Sección 3.1.3, cada elemento del modelo es aplicable a determinados tipos de datos (p. ej. alfanumérico), tipos de colecciones de datos (p. ej. base de datos relacional), dominios de aplicación (p. ej. salud) y usos de datos (p. ej. toma de decisiones). Se utiliza la siguiente notación para cada elemento del modelo que sea aplicable solamente a un subconjunto de tipos de datos (TD), tipos de colección (TC), dominios de aplicación (DA) o usos de datos (UD):

ElementoModelo [TD:td_1,...,td_w; TC:tc_1,...,tc_x; DA:da_1,...,da_y; UD:ud_1,...,ud_z]

Por ejemplo, si un elemento del modelo se anota como *Consistencia Topológica (TD:«Geo»)*, esto indica que la Consistencia Topológica (factor) es únicamente aplicable al tipo de datos indicado (i.e. tipo de dato «Geo»: datos geográficos). A su vez, como no se explicitan, el elemento es aplicable a cualquier tipo de colección, dominio de aplicación y uso de datos.

3.3. Dimensión Exactitud

Esta sección describe la dimensión Exactitud y sus factores: Correctitud Semántica, Correctitud Sintáctica, Precisión, Exactitud Posicional Absoluta y Exactitud Posicional Relativa. Estos dos últimos factores se utilizan exclusivamente para el tipo de datos geográfico. La Tabla 3.5 presenta los datos generales de esta dimensión.

Tabla 3.5: Dimensión Exactitud

Exactitud	
Definición	Proximidad entre un valor de datos *v* y un valor de datos *v'*, considerado como la representación correcta del fenómeno del mundo real que *v* intenta representar. (Adaptado de [BS16])
Otros nombres	Accuracy (ingl.), correctitud
Factores	Correctitud Semántica Correctitud Sintáctica Precisión Exactitud Posicional Absoluta (TD:«Geo») Exactitud Posicional Relativa (TD:«Geo»)

La Tabla 3.6 presenta en forma resumida los datos generales del factor Correctitud Semántica. El detalle de las métricas de este factor se presenta en la Sección 16.1.

Tabla 3.6: Factor Correctitud Semántica

Correctitud Semántica	
Definición	Proximidad entre el valor *v* de un atributo y su verdadero valor *v'*. (Adaptado de [BS16])
Otros nombres	Semantic Correctness (ingl.)
Métricas Genéricas	**CorrectitudSemDebil**: Evalúa si una instancia de un atributo, que no forma parte de la identificación de la entidad a la que pertenece, existe dentro de un referencial de valores posibles de ese atributo. **CorrectitudSemFuerte**: Evalúa si una instancia de un atributo, que forma parte de la identificación de la entidad a la que pertenece, existe dentro de un referencial de valores posibles de ese atributo. **RatioCorrectitudSemFuerte**: Métrica agregada de tipo Ratio basada en CorrectitudSemFuerte. **RatioCorrectitudSemDébil**: Métrica agregada de tipo Ratio basada en CorrectitudSemDébil.

El Ejemplo 4 presenta un problema de Correctitud Semántica.

Ejemplo 4

La Dirección Impositiva y el Instituto de Seguridad Social realizan un cruzamiento de datos entre declaraciones juradas de Impuesto a la Renta y Seguro de Salud ingresadas mediante sus respectivos sistemas Web. De dicho cruzamiento surge que un número significativo de contribuyentes, identificados por su número de Registro Tributario, tienen un código 1 en el campo Código de Salud en uno de los sistemas, en tanto que en el otro sistema tienen diferentes códigos. A partir de este análisis se comienza a investigar si el código 1 en uno de los sistemas puede haber sido ingresado por error (p. ej. por ser el valor por defecto) quedando esos contribuyentes con un seguro de salud incorrecto.

La Tabla 3.7 presenta en forma resumida los datos generales del factor Correctitud Sintáctica. El detalle de las métricas de este factor se presenta en la Sección 16.1.

Tabla 3.7: Factor Correctitud Sintáctica

Correctitud Sintáctica	
Definición	Proximidad entre el valor *v* de un atributo y los elementos del dominio de definición de dicho atributo. (Adaptado de [BS16])
Otros nombres	Syntactic Correctness (ingl.)
Métricas Genéricas	**Formato**: Indica si el valor de un atributo cumple con el formato definido para ese atributo según algún estándar o diccionario.
Métricas Específicas	**Formato(Pais, ISOAlpha3)**: Indica si el código de un país está en el formato ISO 3166-1 alpha-3 (3 letras). **Formato(Enfermedad, CIE10) [DA:«Salud»]**: Indica si el código de una enfermedad está en el formato CIE-10 (comienza con una letra válida y tiene una cantidad menor o igual a 6 dígitos). **Formato(NumeroDocumento, DNIC)**: Indica si el valor del atributo numDoc cumple con el formato de cédula de identidad uruguaya establecido por DNIC, que establece que la cédula tiene siete dígitos seguidos de un octavo dígito verificador que está en función de los otros siete ($d_1d_2d_3d_4d_5d_6d_7 - v, v = f(d_i)$).

El Ejemplo 5 presenta un problema de Correctitud Sintáctica.

Ejemplo 5

Con el objetivo de conocer la composición racial de los habitantes y determinar los perfiles demográficos y socio económicos de la población, el Instituto de Estadística releva el dato conocido como Autopercepción de la Ascendencia Racial. Este dato se construye a través de una serie de cinco preguntas de respuesta sí/no en donde el ciudadano contesta si cree tener ascendencia negra, amarilla, blanca, indígena u otra. Con las respuestas se construye un código binario de cinco dígitos que se guarda como un string (p. ej. «10110»). Se encuentran códigos que no cumplen con esa sintaxis, como «1F», «30», «LF», «ESC», «101000101000», etc.

La Tabla 3.8 presenta en forma resumida los datos generales del factor Precisión. El detalle de las métricas de este factor se presenta en la Sección 16.1.

Tabla 3.8: Factor Precisión

Precisión	
Definición	Captura el grado de detalle que posee un dato que lo hace útil para un determinado uso o que permite discriminarlo de otros datos que no son exactamente iguales. (Adaptado de [BS16]) y [ISO08])
Otros nombres	Precision (ingl.)
Métricas Genéricas	**Escala**: En el caso de valores numéricos, se calcula como $1 - \frac{error}{valor}$n en donde error está dado por el instrumento de medición y valor es el valor del dato **ErrorEstandar**: Desviación estándar de un conjunto de datos

El Ejemplo 6 presenta un problema de Precisión.

Ejemplo 6

En un sistema de trazabilidad de procesos de negocio se registra, entre otros datos, un *timestamp* con la fecha y hora de inicio de un paso de proceso de negocio y otro con la fecha y hora de fin del mismo. Se encuentra que el sistema está guardando esos *timestamps* con el formato «yyyy-MM-dd» en lugar de «yyyy-MM-dd HH:mm:ss». Por este motivo no es posible calcular la duración exacta de un proceso, particularmente en aquellos casos de procesos de corta duración que se completan en horas o minutos, y no en días.

La Tabla 3.9 presenta en forma resumida los datos generales del factor Exactitud Posicional Absoluta. El detalle de las métricas de este factor se presenta en la Sección 16.1.

Tabla 3.9: Factor Exactitud Posicional Absoluta [TD:«Geo»]

Exactitud Posicional Absoluta [TD:«Geo»]	
Definición	Proximidad de los valores reportados de las coordenadas a los valores verdaderos o aceptados como tales [ISO13]
Otros nombres	Absolute positional accuracy (ingl.), Exactitud posicional externa.
Métricas Genéricas	**ÍndiceErroresPosicionalesPorUmbral**: Ratio entre el número de incertidumbres posicionales superiores a un umbral dado para un conjunto de posiciones, sobre número total de las posiciones medidas (Tabla D.33 de [ISO13]). Esta métrica suele ser evaluada a través de muestreos. **ValorMedioIncertidumbrePosicional**: Distancia entre la posición medida y la que se considera como verdadera (Tabla D.29 de [ISO13]). Esta métrica suele ser evaluada a través de muestreos.

La Tabla 3.10 presenta en forma resumida los datos generales del factor Exactitud Posicional Relativa. El detalle de las métricas de este factor se presenta en la Sección 16.1.

Tabla 3.10: Factor Exactitud Posicional Relativa [TD:«Geo»]

Exactitud Posicional Relativa (TD:«Geo»)	
Definición	Proximidad de las posiciones relativas de los objetos geográficos de un conjunto de datos a sus respectivas posiciones relativas verdaderas o aceptadas como tales [ISO13]
Otros nombres	Relative positional accuracy (ingl.), Exactitud posicional interna.
Métricas Genéricas	**ErrorHorizontalRelativo**: Evaluación de los errores aleatorios en la posición horizontal de una entidad geográfica en relación a otra de la misma capa geográfica (Tabla D.55 de [ISO13]).

Es importante mencionar que para las métricas de exactitud posicional es usual recurrir a métodos que sean por lo menos tres veces más precisos que aquello que se desea controlar. Para esto se puede recurrir a: relevamientos directos en campo, imágenes satelitales y aéreas, u otros conjuntos de datos. Es importante que los controles se realicen considerando el universo de discurso (vista del mundo real o hipotético que incluye todo aquello que es de interés [ISO13]) y no la realidad en toda su expresión.

3.4. Dimensión Consistencia

Esta sección describe la dimensión Consistencia y sus factores: Integridad Intra-entidad, Integridad Inter-entidad, Integridad de Dominio y Consistencia Topológica. La Tabla 3.11 presenta los datos generales.

Tabla 3.11: Dimensión Consistencia

Consistencia	
Definición	Captura la violación de las reglas semánticas definidas sobre un conjunto de entidades de negocio o de sus atributos. En un modelo relacional, las restricciones de integridad son un ejemplo de tales reglas semánticas. (Adaptado de [BS16])
Otros nombres	Consistency (ingl.), cohesión, coherencia
Factores	Integridad Inter-entidad Integridad Intra-entidad Integridad de Dominio Consistencia Topológica (TD:«Geo»)

La Tabla 3.12 presenta en forma resumida los datos generales del factor Integridad Inter-entidad. El detalle de las métricas de este factor se presenta en la Sección 16.2.

Tabla 3.12: Factor Integridad Inter-entidad

Integridad Inter-entidad	
Definición	Captura la satisfacción de reglas entre atributos de diferentes entidades de negocio. (Adaptado de [BS16])
Otros nombres	Referential Integrity (ingl.), Integridad Inter-relacion
Métricas Genéricas	**ReglaIntegridadInterEntidad**: Regla de inclusión (clave foránea) o expresión condicional entre atributos de diferentes entidades. **ReglaEspacial [TD:«Geo»]**: Expresión condicional que verifica la ocurrencia de determinada relación espacial topológica entre dos atributos geométricos de diferentes entidades.
Métricas Específicas	**ReglaIntegridadInterEntidad(Sexo, Enfermedad) [DA:«Salud»]**: Indica si una enfermedad especificada en una entidad de negocio relacionada a una persona (p. ej. historia clínica, certificado de defunción) es compatible con el sexo de esa persona.

El Ejemplo 7 presenta un problema de Integridad Inter-entidad que también se considera de consistencia lógica según [ISO13].

Ejemplo 7

En un determinado conjunto de datos geográficos, existe una capa de puntos que representan Aeropuertos y otra de polígonos que representan Lagos. En la capa de Aeropuertos hay un aeropuerto que está dentro de un lago de la otra capa, lo que no es posible que suceda en la realidad.

La Tabla 3.13 presenta en forma resumida los datos generales del factor Integridad Intra-entidad. El detalle de las métricas de este factor se presenta en la Sección 16.2.

Tabla 3.13: Factor Integridad Intra-entidad

Integridad Intra-entidad	
Definición	Captura la satisfacción de reglas entre atributos de una misma entidad. (Adaptado de [BS16])
Otros nombres	Relation Integrity (ingl.)
Métricas Genéricas	**ReglaIntegridadIntraEntidad**: Reglas de dependencia de clave y unicidad de atributos, de dependencias funcionales o de dependencias de atributos. **RatioIntegridadIntraEntidad**: Porcentaje de datos que satisfacen una métrica de Integridad Intra-entidad.
Métricas Específicas	**ReglaIntegridadIntraEntidad(Sexo,Enfermedad) [DA:«Salud»]**: Indica si una enfermedad de una persona es compatible con el sexo de esa persona.

El Ejemplo 8 y el Ejemplo 9 presentan problemas de Integridad Intra-entidad.

Ejemplo 8

En una tabla llamada Direcciones donde se guardan datos alfanuméricos de direcciones de Uruguay, se encuentra una tupla con atributos nombreCalle=«Coronel Alegre» y nombreDepartamento=«Rocha». El problema consiste en que no existe la calle Coronel Alegre en el departamento de Rocha, por lo tanto hay un error en alguno de los dos atributos.

Ejemplo 9

En una tabla llamada Consultas de la historia clínica de un paciente, se encuentra una tupla con un paciente con los atributos sexo=1 (masculino) y diagnóstico=«N80.0», que corresponde al código CIE-10 de la «Endometriosis de útero», que no puede ocurrir en un paciente de sexo masculino.

La Tabla 3.14 presenta en forma resumida los datos generales del factor *Integridad de Dominio*. El detalle de las métricas de este factor se presenta en la Sección 16.2.

Tabla 3.14: Factor Integridad de Dominio

Integridad de Dominio	
Definición	Captura la satisfacción de reglas sobre los valores posibles que puede tomar un atributo. (Adaptado de [BS16])
Otros nombres	Domain Integrity (ingl.)
Métricas Genéricas	**ValoresPosiblesPorExtensión**: Indica si el valor de un atributo se encuentra dentro de un dominio definido por extensión. **ValoresPosiblesPorComprensión**: Indica si el valor de un atributo se encuentra dentro de un dominio definido por comprensión, el cual puede estar dado por una propiedad que cumplen los elementos de ese dominio o por el tipo de dato conocido de ese dominio.
Métricas Específicas	**ValoresPosiblesPE(Sexo, AGESIC)**: Indica si el valor de un atributo Sexo se encuentra dentro del conjunto de valores definidos en el Vocabulario de Persona de AGESIC ([AGE18]).

El Ejemplo 10 y el Ejemplo 11 presentan problemas de integridad de dominio. El Ejemplo 12 describe un problema de integridad de dominio, que también se considera de consistencia de dominio según [ISO13].

Ejemplo 10

En una tabla CertificadosNacidosVivos en donde se guardan datos de nacimientos, una tupla tiene un atributo semanasGestacion=400. Este valor está fuera del rango establecido entre 26 y 52 semanas, ya que se considera que duraciones menores o mayores son altamente improbables para nacidos vivos.

Ejemplo 11

En una tabla de Personas, que utiliza el Vocabulario de Persona de AGESIC ([AGE18]), una tupla tiene el atributo tipoDocumento=69999, que no existe dentro de los códigos definidos actualmente por UNAOID (el código mayor es el 69096) para los tipos de documento de identificación de una persona (cédula de identidad, pasaporte, pasaporte diplomático, etc).

Ejemplo 12

En una capa geográfica de polígonos que corresponden a Planes de Uso de Suelo (PDU), algunos PDU tienen el atributo cultivoInvierno=«Soja», pero los valores permitidos para cultivos de invierno son: Barbecho, Cebada, Cereales de invierno, Colza, Cultivo de cobertura, Pastura Consociada, Pastura no consociada, Pasturas y Trigo.

La Tabla 3.15 presenta en forma resumida los datos generales del factor Consistencia Topológica. El detalle de las métricas de este factor se presenta en la Sección 16.2.

Tabla 3.15: Factor Consistencia Topológica

Consistencia Topológica [TD:«Geo»]	
Definición	Corrección de las características topológicas codificadas explícitamente. Las características topológicas de un conjunto de datos describen las relaciones geométricas entre los ítems del conjunto de datos que no son alteradas por transformaciones elásticas (rubber-sheet transformations) [ISO13]
Otros nombres	Topological Consistency (ingl.)
Métricas Genéricas	**ÍndiceFallosConexiónNodosEnlace**: Porcentaje de fallos en las conexiones de nodos de enlace del total de conexiones de nodos de enlace, medido sobre la geometría de una entidad geográfica (Adaptado de [ISO13] [IDE18]).

El Ejemplo 13 describe un problema de consistencia topológica.

Ejemplo 13

En una capa geográfica de líneas que corresponden a ejes de calles de Montevideo, el eje de Ejido hacia el norte de Isla de Flores tiene un nodo de enlace con el eje de Isla de Flores, y el eje hacia el sur tiene otro nodo de enlace, cuando los dos ejes de Ejido deberían llegar al mismo nodo de enlace sobre la calle Isla de Flores.

3.5. Dimensión Completitud

Esta sección describe la dimensión Completitud y sus factores: Cobertura, Densidad y Comisión. La Tabla 3.16 presenta los datos generales de la dimensión.

Tabla 3.16: Dimensión Completitud

Completitud	
Definición	Captura la medida en que los datos son de la amplitud, profundidad y alcance suficientes para una determinada tarea. (Adaptado de [BS16])
Otros nombres	Completeness (ingl.)
Factores	Cobertura Densidad Comisión (TD:«Geo»)

La Tabla 3.17 presenta en forma resumida los datos generales del factor Cobertura. El detalle de las métricas de este factor se presenta en la Sección 16.3.

Tabla 3.17: Factor Cobertura

Cobertura	
Definición	Captura la proporción entre la cantidad de entidades existentes en una determinada colección de datos, y el total de entidades que deberían existir en dicha colección. La cobertura varía si se utiliza la Asunción de Mundo Cerrado, según la cual una colección de datos debería contener todas las entidades de un tipo, o si se utiliza la Asunción de Mundo Abierto, según la cual una colección de datos puede ser una representación parcial de las entidades del mundo real. (Adaptado de [BS16])
Otros nombres	Coverage (ingl.)
Métricas Genéricas	**RatioCobertura**: Proporción entre la cantidad de instancias de una entidad y el número total de instancias de un referencial de esa entidad.

El Ejemplo 14 describe un problema de cobertura.

> **Ejemplo 14**
>
> El Ministerio del Interior posee una tabla con los hurtos registrados en 2018. Los números anuales de 2018 se consideran útiles para realizar comparaciones con años anteriores (Asunción de Mundo Cerrado), pero existe información de que muchos hurtos no son denunciados por lo que no figuran en esa tabla (Asunción de Mundo Abierto).

La Tabla 3.18 presenta en forma resumida los datos generales del factor Densidad. El detalle de las métricas de este factor se presenta en la Sección 16.3.

Tabla 3.18: Factor Densidad

Densidad	
Definición	Captura la proporción entre la cantidad de instancias de atributo con valores no nulos y el total de instancias de dicho atributo (Adaptado de [BS16]). Un valor nulo de una instancia de atributo *A* de una entidad *E* puede interpretarse de varias maneras: 1. *E* no posee *A* 2. se desconoce si *E* posee *A* o no 3. *E* posee *A* pero se desconoce su valor
Otros nombres	Density (ingl.)
Métricas Genéricas	**NoNulo**:Indica si una instancia de atributo tiene un valor no nulo. Puede ser necesario especificar un diccionario con todos los valores del atributo que se consideran nulos o vacíos, cuando existe más de una posibilidad. **DensidadPonderada**: Aplica un cálculo sobre algunos atributos de una instancia de entidad, evaluado para cada uno si es nulo o no (como en la métrica NoNulo) pero multiplicando el resultado de cada atributo por un coeficiente entre 0 y 1 cuya suma sea igual a 1. A mayor gravedad de tener un nulo en un atributo, más cercano a 0 será el coeficiente para ese atributo. **RatioNoNulos**: Métrica agregada de tipo Ratio basada en NoNulo. **RatioDensidadPonderada**: Métrica agregada de tipo Ratio basada en DensidadPonderada.

La Tabla 3.19 presenta en forma resumida los datos generales del factor Comisión. El detalle de las métricas de este factor se presenta en la Sección 16.3.

Tabla 3.19: Factor Comisión (TD:«Geo»)

Comisión [TD:«Geo»]	
Definición	Datos excedentes presentes en un conjunto de datos [ISO13]
Otros nombres	Commission (ingl.)
Métricas Genéricas	**ÍtemExcedente**: Indica si una instancia está incorrectamente presente en el conjunto de instancias de una entidad (Tabla D.1 de [ISO13]). **ÍndiceItemsExcedentes**: Número de entidades excedentes en el conjunto total de instancias de una determinada entidad o muestra de datos de esa entidad, en relación al número del total que deberían haber estado presentes (Tabla D.3 de [ISO13]).

3.6. Dimensión Unicidad

Esta sección describe la dimensión *Unicidad* y sus factores: *No-duplicación* y *No-contradicción*. La Tabla 3.20 presenta los datos generales de la dimensión.

Tabla 3.20: Dimensión Unicidad

Unicidad	
Definición	Captura el grado en el que un dato del mundo real es representado en forma única. (Adaptado de [BS16])
Otros nombres	Uniqueness (ingl.). En [BS16] se habla también de *redundancia* para el caso de *linked data*. La norma [ISO13] la considera como parte del factor Compleción y lo establece como una medida para este factor.
Factores	No-duplicación No-contradicción

La Tabla 3.21 presenta en forma resumida los datos generales del factor *no-duplicación*. El detalle de las métricas de este factor se presenta en la Sección 16.4.

Tabla 3.21: Factor No-duplicación

No-duplicación	
Definición	Captura el grado de duplicación (o repetición) de un mismo dato.
Otros nombres	Duplication-free (ingl.)
Métricas Genéricas	**AtributoDuplicado** ***Semántica:*** Indica si una instancia de atributo tiene el mismo valor que otra instancia del mismo atributo. **ConjuntoAtributosDuplicado**:Indica si las instancias de un conjunto de atributos de una instancia de entidad, tienen el mismo valor en otra instancia de entidad. **EntidadDuplicada**: Indica si existe, para una instancia de entidad, al menos otra instancia más que representa el mismo objeto del mundo real, con los mismos datos o algún dato faltante. Se debe especificar un conjunto de atributos que permita identificar unívocamente una instancia de entidad. Los demás atributos de la entidad que no pertenezcan a dicho conjunto deben tener los mismos valores en las dos instancias, o ser nulos en alguna de ellas, para que se consideren duplicados exactos. **RatioAtributoDuplicado**: Métrica agregada de tipo Ratio basada en AtributoDuplicado. **RatioConjuntoAtributosDuplicado**: Métrica agregada de tipo Ratio basada en ConjuntoAtributosDuplicado. **RatioEntidadesDuplicadas**: Métrica agregada de tipo Ratio basada en EntidadDuplicada.

El Ejemplo 15 y el Ejemplo 16 presentan problemas de duplicación.

Ejemplo 15

En un sistema de Historia Clínica Digital, se intercambian mensajes HL7 ADT (Admit Discharge Transfer) que contienen información demográfica del paciente. En cada mensaje, además del identificador principal del paciente, opcionalmente se puede incluir un identificador alternativo o *Alternate Patient ID*. Con el fin de tener datos estadísticos de la cantidad de pacientes que son atendidos por día, otro sistema guarda una tabla diaria con un único registro por paciente atendido cada día. Se detecta que en esa tabla, aparece más de un registro con el mismo *Alternate Patient ID*.

Ejemplo 16

En un sistema de Historia Clínica Digital, se está estudiando la implementación de un Enterprise Master Patient Index (EMPI), para poder identificar datos heterogéneos que corresponden al mismo paciente en distintos sistemas y así mantener un índice único de pacientes. Se quiere dimensionar el problema y justificar la inversión, realizando un estudio de datos duplicados, buscando la repetición exacta de diferentes combinaciones de atributos: a - (primer nombre, primer apellido, fecha de nacimiento), b - (primer nombre, segundo nombre, primer apellido, segundo apellido, fecha de nacimiento), c - (primer nombre, primer apellido, fecha de nacimiento, sexo). Los resultados muestran la existencia de duplicados: a - 20 %, b - 16 %, c - 14 %.

La Tabla 3.22 presenta en forma resumida los datos generales del factor No-contradicción. El detalle de las métricas de este factor se presenta en la Sección 16.4.

Tabla 3.22: Factor No-contradicción

No-contradicción	
Definición	Captura el grado de duplicación (o repetición) de una misma instancia de entidad del mundo real que es representada con datos contradictorios.
Otros nombres	Contradiction-free (ingl.)
Métricas Genéricas	**EntidadContradictoria**: Indica si existe, para una instancia de entidad, al menos otra instancia más que representa el mismo objeto del mundo real, con alguna contradicción entre sus datos. Dado que las dos instancias de entidad pueden tener distinta clave, se debe especificar una función de similitud que compare un conjunto de atributos en ambas para detectar si es la misma entidad (técnica de resolución de entidades). La función de similitud puede utilizar la distancia de Levenshtein, la correspondencia de trigramas y algoritmos fonéticos como Soundex, Metaphone, etc. (Capítulo 8 de [BS16]) para detectar las entidades iguales dentro de un cierto umbral. En el caso que las dos entidades tengan los mismos valores en ese conjunto de atributos, sólo se considerarán contradictorias si tienen valores diferentes en los demás atributos, ya que en caso contrario serían entidades duplicadas y no contradictorias. **RatioEntidadContradictoria**: Métrica agregada de tipo Ratio basada en EntidadContradictoria.

El Ejemplo 17 presenta problemas de contradicción.

Ejemplo 17

En una empresa de telecomunicaciones se está trabajando en la unificación de sus sistemas de telefonía fija y móvil, para lo que se generó una única tabla de clientes en base a su documento de identidad. Existen dudas de la confiabilidad de ese dato en los sistemas de origen, ya que se encontraron entidades duplicadas con contradicciones como las siguientes:

1. c1=(documento:2837451-6, nombre:«Mariana», apellido:«Rinaldi», direccion:«Comandante Braga 2060», telefono: 099123321, profesion:«estudiante»)
2. c2=(documento:2837451-3, nombre«María», apellido:«Rinaldi», direccion:«Comandante Braga 2060», telefono: 098116622, profesion:«bailarina»)

3.7. Dimensión Frescura

Esta sección describe la dimensión *Frescura* y sus factores: *Actualidad* y *Oportunidad*. La Tabla 3.23 presenta los datos generales de la dimensión.

Tabla 3.23: Dimensión Frescura

Frescura	
Definición	Captura la rapidez con la que los cambios en el mundo real son reflejados en la actualización de los datos. La frescura es un tipo de exactitud no-estructural dependiente de la variable tiempo, lo que implica que un dato que un determinado momento es correcto, puede no serlo en otro momento. (Adaptado de [BS16])
Otros nombres	Freshness (ingl.), exactitud temporal
Factores	Actualidad Oportunidad

La Tabla 3.24 presenta en forma resumida los datos generales del factor *Actualidad*. El detalle de las métricas de este factor se presenta en la Sección 16.5.

Tabla 3.24: Factor Actualidad

Actualidad	
Definición	Captura el tiempo de demora entre un cambio en el mundo real y la correspondiente actualización de los datos.
Otros nombres	Currency (ingl.)
Métricas Genéricas	**DesactualizaciónPorFecha**: Sea t_a la fecha de acceso a un dato (por defecto es la fecha actual), t_w la fecha del último cambio del dato en el mundo real (o en otra colección de datos de referencia) y t_u la fecha de la última actualización del dato en la colección objetivo. Si $t_u \geqslant t_w$, la métrica devuelve 0. Si $t_u < t_w$, la métrica devuelve la diferencia t_u - t_w. **DesactualizaciónPorCambios**: Indica la cantidad de cambios sufridos por un dato en el mundo real (o en otra colección de datos de referencia) luego de la última actualización del dato en la colección objetivo. **DesactualizaciónPorFormato**: Chequea si un dato se encuentra desactualizado en base a reglas sintácticas que establecen cuando un dato es actual y cuando no, en base al formato vigente de su tipo de dato.
Métricas Específicas	**DesactualizaciónPorFormato(TelefonoFijo, FormatoPNN**: Chequea si un número de telefonía fija se encuentra desactualizado en base al formato vigente del Plan Nacional de Numeración (PNN), que establece que a partir de 2010, todos los teléfono pasaron a tener 8 dígitos.

El Ejemplo 18 presenta problemas de actualidad.

Ejemplo 18

En un sistema donde se registran periódicamente las consultas médicas de niños, se encuentra que la estatura de algunos niños no cambió en sucesivas consultas mensuales, por lo que se determina que esos valores están desactualizados.

La Tabla 3.25 presenta en forma resumida los datos generales del factor *Oportunidad*. El detalle de las métricas de este factor se presenta en la Sección 16.5.

Tabla 3.25: Factor Oportunidad

Oportunidad	
Definición	Captura la demora que existe entre la actualización de un dato y el momento en el que éste se encuentra disponible para ser utilizado.
Otros nombres	Timeliness (ingl.)
Métricas Genéricas	**BoolOportunidadAtributoPorFecha**: Indica si el valor actualizado de una instancia de atributo está disponible antes de una fecha límite. **BoolOportunidadAtributoPorIntervalo**: Indica si el valor actualizado de una instancia de atributo está disponible dentro de un intervalo de vigencia. **BoolOportunidadEntPorFecha**: Indica si una instancia de entidad con sus datos actualizados está disponible antes de una fecha límite. La fecha límite puede ser única o dependiente de cada instancia de la entidad. **BoolOportunidadEntPorIntervalo**: Indica si una instancia de entidad está disponible con sus datos actualizados dentro de un intervalo de vigencia. El intervalo de vigencia puede ser único o dependiente de cada instancia de la entidad.

El Ejemplo 19 presenta problemas de oportunidad.

Ejemplo 19

En un sistema de emisión de alertas meteorológicas, se envían mensajes hacia todos los suscriptores advirtiendo sobre situaciones climáticas peligrosas, con una hora de inicio y una hora de fin de la alerta. Se constata que muchos mensajes llegan a los suscriptores en un tiempo posterior a la hora de inicio de la alerta.

4 Caso de Estudio

Este capítulo presenta un caso de estudio que plantea una escenario posible para ejemplificar el uso del *framework*. La Sección 4.1 presenta una descripción general del caso de estudio. La Sección 4.2 y la Sección 4.3 brindan detalles del escenario planteado. La Sección 4.4 presenta el modelo conceptual de los datos del escenario. La Sección 4.5 describe aspectos de calidad de datos considerados en el caso de estudio.

4.1. Descripción General

El caso de estudio plantea un sistema denominado Sistema de Reclamos Ciudadanos (SRC), que está inspirado en el Sistema Único de Respuesta (SUR[1]) de la aplicación móvil de la Intendencia de Montevideo (IM). La Figura 4.1 presenta una visión conceptual del sistema SRC y los distintos tipos de usuarios involucrados.

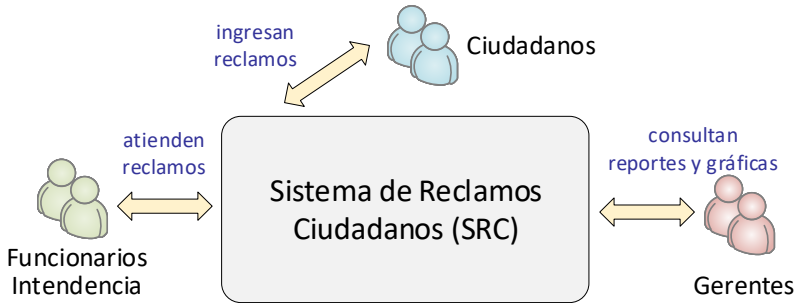


Figura 4.1: Sistema de Reclamos Ciudadanos (SRC)

[1] http://montevideo.gub.uy/areas-tematicas/servicios-digitales/sistema-unico-de-respuesta

Los **Ciudadanos** son los usuarios finales del sistema SRC y realizan reclamos a través de una aplicación en su teléfono celular o dispositivo móvil.

Los **Funcionarios** son los que procesan y atienden los reclamos de los ciudadanos. Los reclamos se asignan a diferentes áreas de la intendencia de acuerdo a la categoría del reclamo (p. ej. contenedor lleno).

Los **Gerentes**, tanto de intendencias como de otros organismos, tienen acceso a funcionalidades de reportes y gráficas que les brindan soporte para la toma de decisiones (p. ej. planificar mejoras en el sistema).

La Figura 4.2 presenta el diagrama de casos de uso del sistema SRC, donde se pueden observar las funcionalidades a las que tienen acceso los distintos tipos de usuario del sistema.

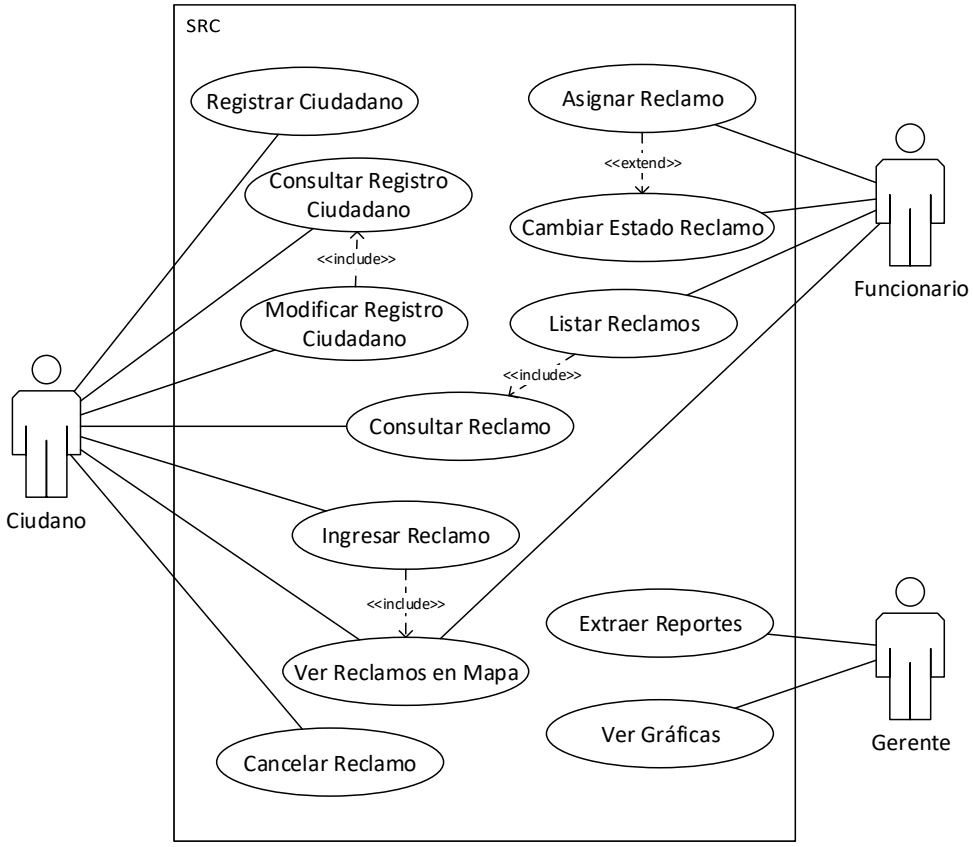


Figura 4.2: Casos de Uso del SRC

A modo de resumen entonces, el sistema SRC brinda soporte para:

1. la **gestión de reclamos**: los ciudadanos reportan reclamos a las intendencias y los funcionarios de las intendencias atienden dichos reclamos
2. la **toma de decisiones**: los gerentes de las intendencias u otros organismos consultan reportes y gráficas para la toma de decisiones

En la Sección 4.2 y en la Sección 4.3 se describe cómo el sistema SRC brinda soporte para la gestión de reclamos y para la toma de decisiones, respectivamente.

4.2. Soporte a la Gestión de Reclamos

La Figura 4.3 presenta los principales componentes del sistema SRC vinculados a la gestión de reclamos en una intendencia específica: la Intendencia de Montevideo (IM).

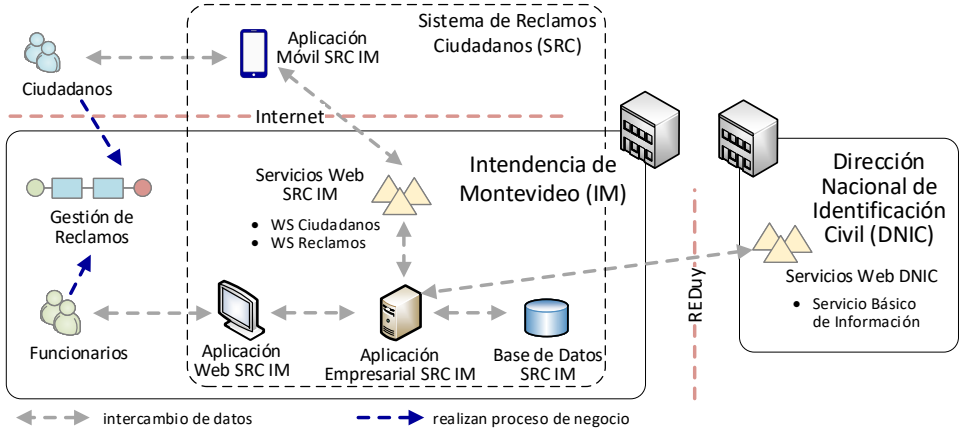


Figura 4.3: Componentes del Sistema SRC para la Gestión de Reclamos

Los ciudadanos pueden registrarse en el sistema y realizar reclamos a través de una aplicación para dispositivos móviles. Para realizar un reclamo, el ciudadano debe especificar la categoría del reclamo y agregar un marcador en un mapa con la ubicación aproximada del mismo, así como una descripción y opcionalmente una o varias fotos. Los reclamos se muestran en el mapa con marcadores cuyo color indica la categoría (p. ej. saneamiento, limpieza, arbolado).

Los reclamos ingresados son procesados por un funcionario de la intendencia a través de una aplicación Web. Los funcionarios pueden asignar los reclamos a un área específica de la intendencia para su resolución, así como rechazarlos en caso de que ya estén resueltos o no sean válidos (p. ej. el reclamo tiene datos insuficientes). En toda operación los funcionarios pueden agregar observaciones. Una vez que un área recibe y resuelve un reclamo, los funcionarios de esa área deben indicar que el reclamo está resuelto cambiando su estado. Los ciudadanos pueden cancelar un reclamo si éste no fue rechazado ni resuelto.

La lógica de negocio del sistema SRC está implementada en una aplicación empresarial, la cual es accedida tanto por la aplicación Web que utilizan los funcionarios así como por la aplicación móvil que utilizan los ciudadanos. En particular, la aplicación móvil accede a la funcionalidad de la aplicación empresarial a través de servicios Web (i.e. WS Ciudadanos, WS Reclamos).

La aplicación empresarial utiliza una base de datos relacional para almacenar los datos que gestiona (p. ej. reclamos, ciudadanos) y consume un servicio Web (i.e. Servicio Básico de Información) de la Dirección Nacional de Identificación Civil (DNIC) para validar el número de cédula de identidad ingresado en el registro de un ciudadano.

El sistema SRC se utiliza también en otras intendencias (p. ej. Intendencia de Canelones, Intendencia de Maldonado) con distintas categorías de reclamos (p. ej. las intendencias de departamentos sin costas no incluyen reclamos relativos a playas).

4.3. Soporte a la Toma de Decisiones

Para brindar soporte a la toma de decisiones, los datos del sistema SRC en las distintas intendencias se vuelcan a un Data Warehouse (DW) central alojado en Presidencia. Como se puede observar en la Figura 4.4, los datos se cargan en el DW utilizando procesos de Extracción, Transformación y Carga (Extract, Transform and Load, ETL) también alojados y gestionados en Presidencia.

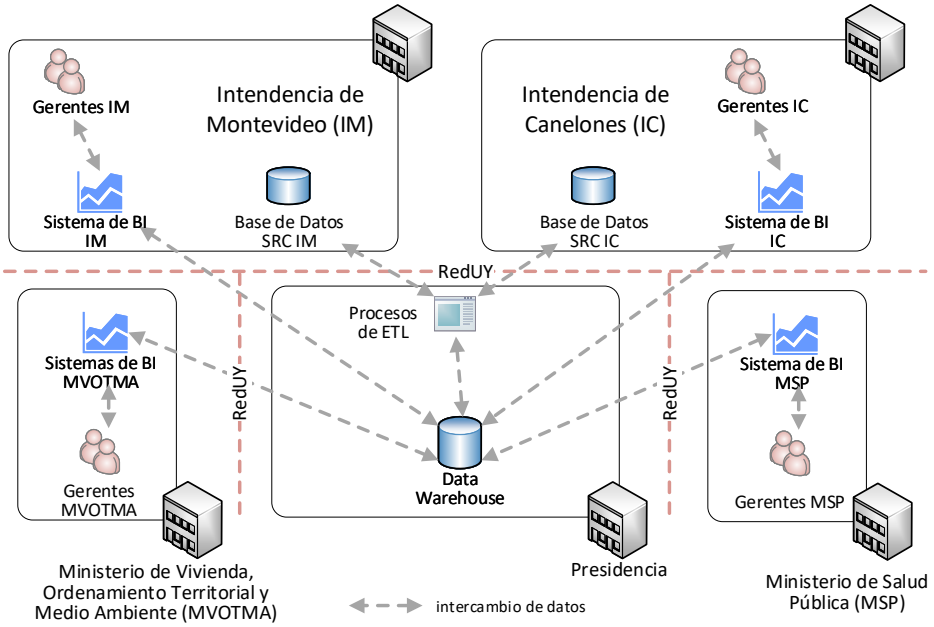


Figura 4.4: Componentes del Sistema SRC para la Toma de Decisiones

El DW permite que los gerentes de las intendencias realicen consultas estadísticas de los datos de los reclamos de su intendencia. Por otro lado, algunas categorías de reclamos son de particular interés para otros organismos. Por ejemplo, los reclamos relativos a playas pueden ser de interés para el Ministerio de Vivienda, Ordenamiento Territorial y Medio Ambiente (MVOTMA[2]) y para el Ministerio de Salud Pública (MSP[3]).

4.4. Modelo de Datos y Datos Referenciales

La Figura 4.5 presenta el modelo conceptual de los datos que maneja el sistema SRC para la gestión de reclamos.

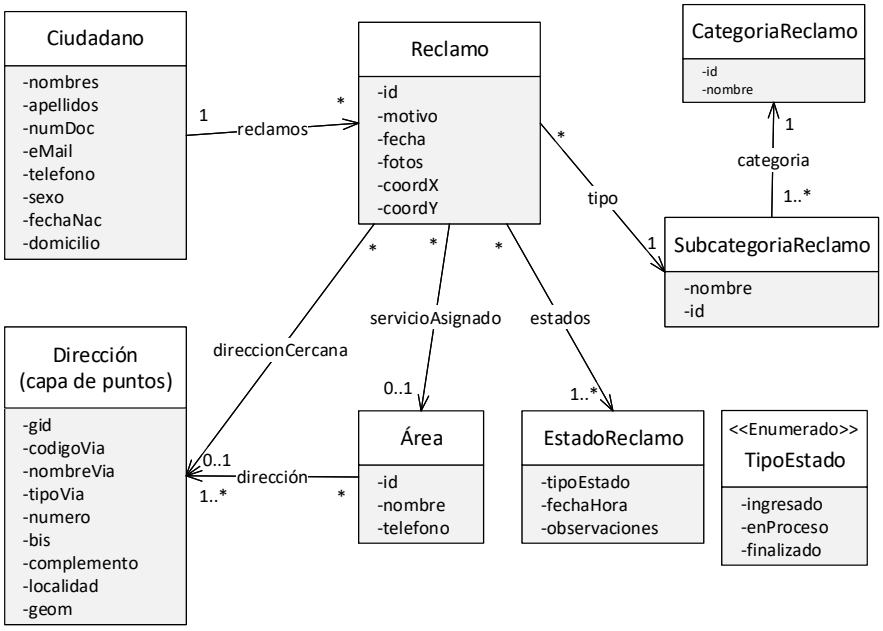


Figura 4.5: Modelo Conceptual de los Datos del SRC

En este modelo se identifican como entidades centrales los ciudadanos y los reclamos con sus estados (p. ej. ingresado, rechazado, resuelto). Se considera además una entidad Dirección que es una capa geográfica con la localización de direcciones y que está vinculada a las entidades reclamo y área.

El modelo también incluye las categorías y subcategorías de los reclamos así como las áreas de la intendencia que atienden los reclamos. Los datos correspondientes a estas entidades dependen de cada intendencia y no cambian mucho en el tiempo, por lo que se cargan cuando se inicializa el sistema SRC.

[2]http://www.mvotma.gub.uy/
[3]https://www.gub.uy/ministerio-salud-publica/

La Tabla 4.1 presenta las categorías y subcategorías de reclamos que maneja la IM en el contexto del caso de estudio.

Tabla 4.1: Categorías y Subcategorías de Reclamos de la IM

Categoría	**Subcategoría**
Limpieza	Contenedor lleno
	Solicitar Contenedor
	Falta de barrido en avenidas y bulevares
	Falta de barrido en calles internas
	Falta de limpieza en rambla
	Falta de limpieza en playa
	Falta de limpieza posterior a feria
Saneamiento	Boca de tormenta obstruida
Calles y veredas	Bache o desnivel
	Bache o desnivel con agua
Arbolado	Árbol deteriorado o riesgoso
	Solicitud de plantación
Alumbrado	Luminaria quemada
Salud	Presencia de cianobacterias en playa
Tránsito	Vehículo estacionado en lugar prohibido

La Tabla 4.2 lista las áreas de atención de reclamos de la IM en el contexto del caso de estudio.

Tabla 4.2: Áreas de Atención de Reclamos de la IM

Alumbrado
Arbolado
Barométrica
Calles y veredas
Centro Coordinador de Emergencias Departamentales (CECOED)
Espacios Públicos
Limpieza
Saneamiento
Tránsito

4.5. Aspectos de Calidad de Datos

Como resultado de la puesta operativa del SRC durante un determinado tiempo, se constata que existen algunos problemas en la calidad de los datos, lo que sirve como motivación inicial para realizar un estudio sistemático y en profundidad, como propone el *framework* que se describe en este documento.

Dentro de los problemas detectados se encuentran los siguientes:

1. **Nombres irreales.** Se han detectado nombres de fantasía, apodos o *nicknames* en los campos destinados a los nombres y apellidos del ciudadano.

2. **Correos electrónicos inexistentes.** A algunos ciudadanos registrados no les llegan los correos electrónicos que envía el sistema porque ingresaron incorrectamente su dirección de correo.

3. **Domicilios no encontrados.** Algunos de los domicilios ingresados por los ciudadanos no se corresponden con ninguna dirección oficial de la capa de direcciones que maneja el sistema.

4. **Edades poco confiables.** Algunas edades de los ciudadanos registrados (que se calculan en base a la fecha de nacimiento ingresada) resultan poco confiables, por estar fuera de los rangos en que deberían encontrarse los usuarios de esta aplicación (mayores de 10 años y menores de 100 años).

5. **Uso abusivo.** Se han constatado varios casos de usuarios que realizan un uso abusivo del sistema, que incluyen: reclamos de incidentes falsos, múltiples reclamos del mismo usuario sobre el mismo incidente, observaciones o fotos de los reclamos con contenido inapropiado o irrelevante.

6. **Reclamos duplicados.** Una parte importante del trabajo del funcionario que recibe los reclamos es poder identificar los reclamos de distintos ciudadanos que hacen referencia al mismo problema. Si estos reclamos duplicados no son detectados oportunamente, puede suceder que las áreas reciban reclamos que ya fueron resueltos en base a otros reclamos.

7. **Reclamos rechazados sin aclaraciones.** Algunos funcionarios no completan las observaciones cuando cambian el estado del reclamo. Esto genera disconformidad en algunos ciudadanos, que ven sus reclamos en estado «rechazado» y no conocen el motivo (p. ej. podría haberse descartado por tratarse de un reclamo duplicado).

8. **Inconsistencia entre la categoría del reclamo y su ubicación geográfica.** Dado que todos los reclamos están georreferenciados, se ha constado que a veces la categoría del reclamo no es compatible con su ubicación. Por ejemplo:

    - Reclamo de la subcategoría «presencia de cianobacterias en playa» es reportado en una ubicación alejada de la playa.

    - Reclamo de la subcategoría «falta de barrido en avenidas y bulevares» es reportado en una calle que no es ni avenida ni bulevar.

9. **Inconsistencia entre la categoría del reclamo y su fecha** En algunos casos, la fecha en la que se ingresa el reclamo no es compatible con la categoría del reclamo. Por ejemplo, se ingresa un reclamo de la subcategoría «falta de limpieza posterior a feria» en una fecha muy posterior a la realización de dicha feria.

5 Proceso para Gestión de Calidad

Este capítulo presenta una descripción general del proceso para la gestión de la calidad de datos del *framework*. El proceso define los roles involucrados y las etapas a seguir para gestionar la calidad de datos en un escenario de gobierno digital.

En la Sección 5.1 se introducen los equipos y roles involucrados en el proceso. En la Sección 5.2 se presentan las principales etapas del proceso, las cuales se refinan en los siguientes capítulos.

5.1. Equipos y Roles del Proceso

Para aplicar el *framework* en un determinado escenario, se requiere la conformación de un Comité de Calidad de Datos (CCD) que será el responsable de llevar a cabo el proceso de gestión de calidad. Esta sección describe los roles que deben integrar este comité, así como otros equipos y roles relacionados.

Cabe recalcar que los equipos y roles que se describen en esta sección deben estar coordinados con otros equipos y roles vinculados a la gestión de datos en una organización. En particular, deben estar en constante coordinación con el Director de Datos de la Organización (Chief Data Officer, CDO) [Lee14], que es el responsable de la gestión y utilización de datos en la organización como un activo organizacional y a menudo estratégico [Bus16].

5.1.1. Roles del Comité de Calidad de Datos

Como se mencionó previamente, un Comité de Calidad de Datos (CCD) es el responsable de llevar a cabo el proceso de gestión de calidad de datos, en un escenario particular. De esta forma, en el marco de una organización se pueden formar distintos CCDs especializados en escenarios particulares.

Un CCD está compuesto por los siguientes roles:

- **Responsable de Calidad de Datos [Int09]:** Responsable de la aplicación del *framework* en un escenario específico. Estará a cargo de todas las etapas del proceso y de dirigir las actividades de los otros integrantes del comité.
- **Analista de Calidad de Datos [Int09]:** Experto en el área de calidad de datos y responsable del análisis de los aspectos de calidad de datos relevantes para el escenario. Algunas de las actividades que están a cargo de este rol son: analizar requerimientos de calidad de datos, identificar actores relevantes para el escenario, identificar problemas de calidad de datos y definir métricas de calidad de datos, entre otros.
- **Técnico de Calidad de Datos:** Experto en el área de calidad de datos con perfil técnico. Responsable de proveer e implementar los recursos técnicos para la aplicación del *framework*. Algunas de las actividades que debe realizar este rol son: examinar datos objetivo e implementar métodos de medición, entre otros.
- **Experto de Negocio [Wen07][Int09]:** Responsable de documentar los requerimientos de negocio y evaluar el impacto que tienen los nuevos requerimientos de negocio sobre la calidad de datos, y viceversa. En el marco del comité, este rol es el referente del área de negocio en la cual se enmarca el escenario.
- **Experto Técnico [Wen07]:** Responsable de la representación de los datos en los sistemas y aplicaciones. En el marco del comité, este rol es el referente de los aspectos técnicos vinculados al escenario.

5.1.2. Otros Equipos y Roles Relacionados

Para avanzar en la gestión de la calidad de datos en una organización, se espera que exista un Comité de Calidad de Datos de la Organización liderado por un Director de Calidad de Datos de la Organización [Wen07]. Este comité también debe estar integrado por expertos de negocio y técnicos, así como (de forma temporal) por un Sponsor de Calidad de Datos.

El Director de Calidad de Datos de la Organización es el responsable de la calidad de los datos en una organización [Wen07]. En particular, este rol es el encargado de designar un CCD para cada escenario en el que se aplique el *framework*.

Por otro lado, un Sponsor de Calidad de Datos es responsable de dirigir, financiar y supervisar la gestión de la calidad de datos [Wen07].

5.2. Principales Etapas del Proceso

El proceso para la gestión de la calidad de datos del *framework* consiste de siete etapas a ser llevadas a cabo por los distintos equipos y roles presentados en la Sección 5.1, las cuales toman como base las etapas descriptas en la Sección 2.3. La Figura 5.1 presenta una visión general de este proceso y sus etapas utilizando BPMN2 [OMG11].

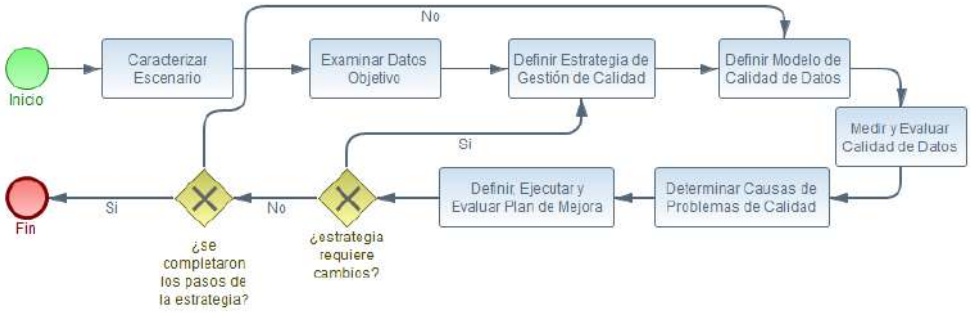


Figura 5.1: Proceso para la Gestión de la Calidad de Datos

La etapa **Caracterizar Escenario** tiene como objetivo identificar los elementos relevantes para el escenario de trabajo, tanto desde un punto de vista técnico y de negocio (p. ej. sistemas que procesan los datos, organizaciones con los que se intercambian datos, procesos de negocio) como desde un punto de vista de la calidad de datos (p. ej. requerimientos de calidad, problemas de calidad). Estos elementos brindan insumos para las siguientes etapas del proceso.

Los principales roles del CCD involucrados en los aspectos técnicos y de negocio de la etapa de caracterización son el Experto de Negocio y el Experto Técnico. Estos aspectos del escenario se describen en el Capítulo 6. Por otro lado, los principales roles del CCD involucrados en los aspectos de calidad de datos de la caracterización son el Experto de Negocio y el Analista de Calidad de Datos. Estos aspectos se describen en el Capítulo 7.

La etapa **Examinar Datos Objetivo** tiene como objetivo conocer las características de los datos y realizar una primera estimación de su calidad, así como detectar problemas de calidad de datos y definir nuevos requerimientos a partir de estos. Una de las principales herramientas de soporte en esta etapa es la aplicación de técnicas de *Data Profiling* [Abe15]. Los principales roles del CCD involucrados en esta etapa son el Experto Técnico, el Experto de Negocio y el Técnico de Calidad de Datos. Esta etapa se describe en el Capítulo 8.

La etapa **Definir Estrategia de Gestión de Calidad** tiene como objetivo definir la estrategia para gestionar la calidad de datos en el escenario de trabajo, en base a pasos que abordan un conjunto de requerimientos de calidad de datos. Los principales roles del CCD involucrados en esta etapa son el Responsable de Calidad de Datos, el Experto de Negocio y el Analista de Calidad de Datos. Esta etapa se describe en el Capítulo 9.

La etapa **Definir Modelo de Calidad de Datos** tiene como objetivo definir el modelo de calidad de datos para el escenario de trabajo. Este modelo se construye en base al modelo de referencia del *framework* presentado en el Capítulo 3 y a los requerimientos de calidad de datos considerados en la estrategia definida en la etapa anterior. Los principales roles del CCD involucrados en esta etapa son el Responsable de Calidad de Datos, el Experto de Negocio y el Analista de Calidad de Datos. Esta etapa se describe en el Capítulo 10.

La etapa **Medir y Evaluar Calidad de Datos** tiene como objetivo realizar mediciones utilizando las métricas y métodos definidos en el modelo de calidad, así como evaluar la calidad de los datos en base a los perfiles y reglas de evaluación, también definidos en el modelo. Los principales roles del CCD involucrados en esta etapa son el Responsable de Calidad de Datos, el Técnico de Calidad de Datos y el Analista de Calidad de Datos. Esta etapa se describe en el Capítulo 11.

La etapa **Determinar Causas de Problemas de Calidad** tiene como objetivo determinar las causas de los problemas de calidad que resulten de las mediciones y evaluaciones realizadas en la etapa anterior. Para esto se pueden tomar como insumo tanto los resultados de las mediciones y evaluaciones, así como los resultados obtenidos de la caracterización del escenario. Los principales roles del CCD involucrados en esta etapa son el Responsable de Calidad de Datos, el Técnico de Calidad de Datos, el Analista de Calidad de Datos, el Experto de Negocio y el Experto Técnico. Esta etapa se describe en el Capítulo 12.

Por último, la etapa **Definir, Ejecutar y Evaluar Plan de Mejora** tiene como objetivo la definición de un plan de mejora para abordar los problemas de calidad detectados, así como la ejecución y posterior evaluación de este plan. Los principales roles del CCD involucrados en esta etapa son el Responsable de Calidad de Datos, el Técnico de Calidad de Datos y el Analista de Calidad de Datos. Esta etapa se describe en el Capítulo 13.

Una vez finalizada la última etapa del proceso, se debe evaluar si la estrategia definida requiere cambios, por ejemplo, en base a si surgieron nuevos problemas o requerimientos que no habían sido identificados inicialmente. Si la estrategia no requiere cambios, se debe continuar con el siguiente paso de la misma hasta completar todo lo que define.

Cabe recalcar también que el proceso que se presenta en la Figura 5.1 debe llevarse a cabo periódicamente de forma tal de poder considerar nuevos elementos del escenario de trabajo (p. ej. requerimientos de calidad) así como para realizar el monitoreo continuo de la calidad de los datos en el escenario.

La Tabla 5.1 presenta un resumen de las actividades y roles involucrados en cada una de ellas.

Tabla 5.1: Principales Roles Involucrados en las Etapas del Proceso

Etapa	Responsable de Calidad de Datos	Analista de Calidad de Datos	Técnico de Calidad de Datos	Experto de Negocio	Experto Técnico
Caracterización Técnica y de Negocio del Escenario				X	X
Caracterización de Calidad del Escenario		X		X	
Examinar Datos Objetivo			X	X	X
Definir Estrategia de Gestión de Calidad	X	X		X	
Definir Modelo de Calidad de Datos	X	X		X	
Medir y Evaluar Calidad de Datos	X	X	X		
Determinar Causas de Problemas de Calidad	X	X	X	X	X
Definir, Ejecutar y Evaluar Plan de Mejora	X	X	X		

6 Caracterización Técnica y de Negocio

La primera etapa del proceso de gestión de calidad de datos consiste en caracterizar el escenario sobre el cual se quiere aplicar el *framework*. El objetivo de esta etapa es identificar los elementos relevantes para el escenario, tanto desde un punto de vista técnico y de negocio como desde un punto de vista de la calidad de datos.

Este capítulo describe los aspectos técnicos y de negocio de la caracterización del escenario. Los aspectos de calidad de datos de la caracterización se describen en el Capítulo 7.

La Sección 6.1 presenta los objetivos y resultados esperados de la caracterización técnica y de negocio. La Sección 6.2 describe el marco conceptual asociado a esta caracterización. La Sección 6.3 detalla las actividades a realizar en la etapa y la Sección 6.4 describe estas actividades en el marco del caso de estudio presentado en el Capítulo 4.

6.1. Objetivos y Resultados Esperados

El objetivo de esta etapa es identificar los elementos relevantes para el escenario de trabajo desde un punto de vista técnico y de negocio, determinando cómo los distintos artefactos tecnológicos (p. ej. aplicaciones, sistemas) brindan soporte a las funcionalidades de negocio del escenario.

En esta etapa se deben identificar, por ejemplo, aplicaciones que procesan datos, organizaciones con las que se intercambian datos, roles de usuario que utilizan las aplicaciones, y colecciones de datos (p. ej. bases de datos) que utilizan estas aplicaciones. Se deben identificar también los procesos y entidades de negocio relevantes para el escenario, así como la forma en que las aplicaciones dan soporte a estos procesos y gestionan las entidades en las colecciones identificadas. Asimismo, se deben identificar los actores de datos relevantes y su vinculación con los distintos elementos del escenario.

La caracterización técnica y de negocio del escenario permite, por un lado, contar con una visión general de los elementos técnicos y de negocio del escenario, así como conocer la forma en que los elementos técnicos brindan soporte a las funcionalidades de negocio. Por otro lado, esta caracterización es un insumo importante para etapas posteriores del proceso.

A modo de ejemplo, la caracterización puede ser utilizada para determinar las causas de problemas de calidad (p. ej. el hecho de que los datos de una entidad no sean completos en un repositorio y se ingresen de forma manual en una aplicación, puede ser un indicio de que faltan realizar controles en dicha aplicación).

Los principales roles del CCD involucrados en esta etapa de caracterización técnica son el Experto de Negocio y el Experto Técnico.

Los resultados esperados de esta etapa son:

- representación gráfica del escenario
- elementos de negocio del escenario (p. ej. procesos, entidades, actores)
- elementos técnicos del escenario (p. ej. aplicaciones, colecciones de datos)
- relaciones entre clientes de datos (p. ej. aplicación consume servicio Web)
- relaciones entre organizaciones y clientes de datos / colecciones de datos (p. ej. organización es responsable de aplicación)
- relaciones entre clientes de datos y colecciones de datos (p. ej. aplicación actualiza base de datos)
- relaciones entre actores de datos y elementos del escenario (p. ej. actor tiene interés en colección)

6.2. Marco Conceptual Asociado

Esta sección describe los principales conceptos asociados a la caracterización técnica y de negocio del escenario. Estos conceptos permiten y facilitan describir aspectos técnicos así como de negocio de los escenarios en los que se puede aplicar el *framework*.

En particular, estos conceptos permiten especificar las organizaciones, entidades de negocio (p. ej. ciudadano), procesos de negocio, actores de datos, colecciones de datos (p. ej. una base de datos relacional), clientes de datos (p. ej. aplicaciones, sistemas), dominios en los que operan las organizaciones (p. ej. salud, energía), el uso que se le da a los datos (p. ej. soporte a la operativa) y qué tipo de operaciones (p. ej. alta, baja) realizan los clientes sobre las colecciones de datos.

Como se comentó previamente, esta información permite tener una visión general de los elementos técnicos así como de negocio del escenario, y es de suma importancia para las etapas posteriores del proceso (p. ej. para la identificación de causas de problemas de calidad de datos).

6.2.1. Organizaciones, Colecciones de Datos y Clientes de Datos

La Figura 6.1 presenta los conceptos del escenario de trabajo relacionados a organizaciones, colecciones de datos y clientes de datos.

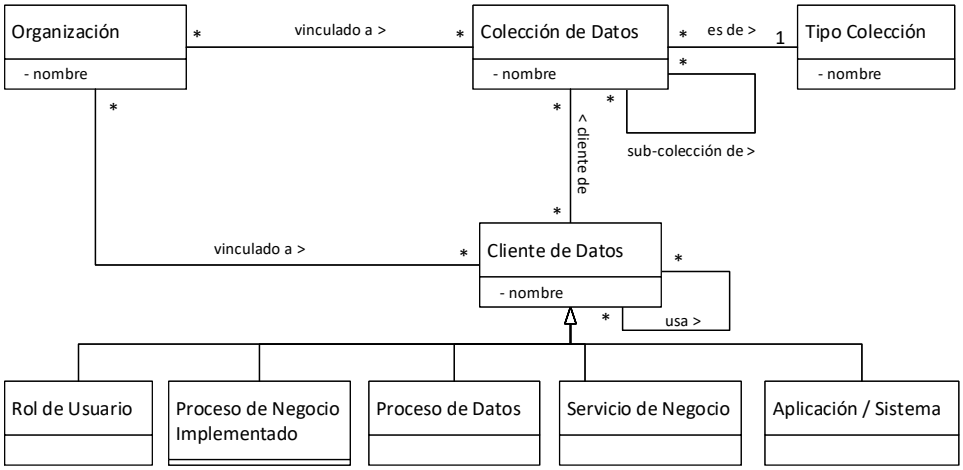


Figura 6.1: Organizaciones, Colecciones de Datos y Clientes de Datos

Las Organizaciones refieren a todos las organizaciones involucradas en un escenario de trabajo. Las organizaciones se vinculan a Colecciones de Datos, por ejemplo, porque son responsables de los datos o porque modifican / consultan los datos.

Las Colecciones de Datos refieren a conjuntos de datos que pueden estar almacenados de distinta forma (p. ej. en una base de datos relacional) o siendo transferidos de un cliente de datos a otro (p. ej. a través de una plataforma informática). Una colección de datos se puede organizar en varias subcolecciones de datos (p. ej. una base de datos relacional se puede organizar en varios esquemas).

Los Tipos de Colección refieren a la forma en que las colecciones de datos se almacenan o acceden (p. ej. en una base de datos relacional). La Tabla 6.1 presenta un conjunto inicial de tipos de colección considerados en el *framework*. Este conjunto puede ser extendido de acuerdo a nuevos requerimientos.

Tabla 6.1: Tipos de Colección

Nombre	**Descripción**
Base de Datos Relacional	Colección de datos almacenada en una base de datos relacional.
Base de Datos Documental	Colección de datos almacenada en una base de datos documental.
Flujo de Datos	Colección de datos que se transfiere de un cliente de datos a otro.

Los Clientes de Datos refieren a distintos tipos de entidades que realizan operaciones sobre las colecciones de datos (p. ej. una aplicación, un proceso de negocio, un rol de usuario). En el marco de un escenario de gestión de calidad de datos, los clientes de datos se vinculan con organizaciones de distintas formas (p. ej. una organización puede ser responsable de un cliente de datos, una organización puede utilizar un cliente de datos).

El Ejemplo 20 presenta parte de la descripción de un escenario utilizando los conceptos descriptos, en el marco del proceso de negocio que permite la generación de certificados de nacido vivo en Uruguay.

Ejemplo 20

Organizaciones:	Ministerio de Salud Pública (MSP) Dirección Nacional de Identificación Civil (DNIC)
Colecciones de Datos:	Certificados de Nacido Vivo (CNV). Tipo: Base de Datos Relacional. Datos Básicos de Ciudadanos (DBC). Tipo: Flujo de Datos.
Clientes de Datos:	**Aplicación**: Certificado de Nacido Vivo Electrónico (apCNVE) **Servicio**: Servicio Básico de Información (servSBI) **Roles de Usuario:** Médico, Administrador
Relaciones:	1. Organización MSP vinculada a servicio servBI 2. Organización MSP vinculada a aplicación apCNVE 3. Organización DNIC vinculada a servicio servBI 4. Rol Médico usa aplicación apCNVE 5. Rol Administrador usa aplicación apCNVE 6. Aplicación apCNVE es cliente de colección CNV 7. Servicio servSBI es cliente de colección DBC

Dos de las organizaciones que participan en el escenario son el Ministerio de Salud Pública (MSP) y la Dirección Nacional de Identificación Civil (DNIC). Dentro de las colecciones de datos involucradas en el escenario hay una base de datos relacional, que almacena información de los certificados de nacido vivo generados, y un flujo de datos que se origina en la invocación de un servicio provisto por la DNIC (i.e. Servicio Básico de Información).

Por otro lado, algunos de los clientes de datos del escenario son la aplicación del MSP para generar estos certificados (i.e. apCNVE), el Servicio Básico de Información (i.e. servSBI) de la DNIC y dos roles de usuario que utilizan la aplicación apCNVE (i.e. Médico, Administrador).

6.2.2. Tipos de Vinculación, Usos de Datos y Dominios de Aplicación

La Figura 6.2 presenta los conceptos del escenario de trabajo relacionados a los tipos de vinculación, usos de datos y dominios de aplicación[1].

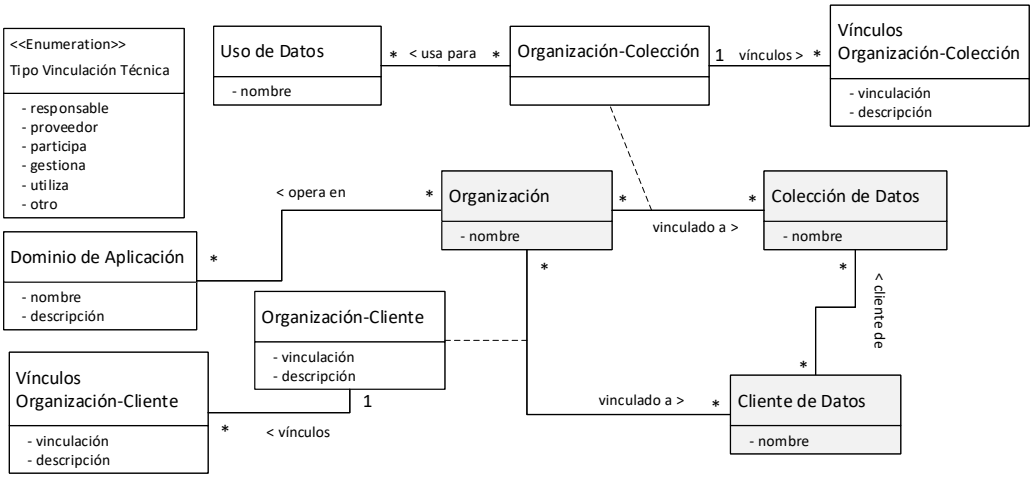


Figura 6.2: Tipos de Vinculación, Usos de Datos y Dominios de Aplicación

Como se mencionó en la Sección 6.2.1 las organizaciones se pueden vincular a colecciones de datos y clientes de datos de distintas formas (p. ej. ser responsables, utilizarlos). La Tabla 6.2 presenta los tipos de vinculación que considera el *framework*.

Tabla 6.2: Tipos de Vinculación de Organizaciones con Clientes y Colecciones de Datos

Nombre	**Descripción**
Responsable	La organización (p. ej. organismo público) tiene la responsabilidad del funcionamiento y/o gestión de la colección o el cliente de datos.
Proveedor	La organización (p. ej. empresa de *Software*) está a cargo del desarrollo, operación o alojamiento de la colección o el cliente de datos.
Participa	La organización participa en la gestión y/o funcionamiento de la colección de datos (p. ej. definiendo estructuras de datos) o cliente de datos (p. ej. definiendo funcionalidades).
Gestiona	La organización gestiona la colección de datos (p. ej. actualizando sus datos) o cliente de datos (p. ej. destinando recursos para su operación).
Utiliza	La organización utiliza la colección de datos (p. ej. consultando datos) o cliente de datos (p. ej. ejecutando funcionalidades).
Otro	La organización tiene otro tipo de vínculo técnico con la colección o cliente de datos.

[1] Los conceptos sombreados son los ya presentados previamente.

Los Dominios de Aplicación refieren a contextos específicos en donde operan las organizaciones. Estos contextos se pueden caracterizar por el sector (p. ej. gobierno, salud) y tipos de actividades (p. ej. cadena de suministro), entre otros. La Tabla 6.3 presenta un conjunto inicial de dominios de aplicación a ser considerados en el *framework*. Este conjunto puede ser extendido de acuerdo a nuevos requerimientos.

Tabla 6.3: Dominios de Aplicación

Nombre
Ciudadanía
Salud
Energía
Gobierno
Ciudades Inteligentes
Finanzas
Energía
Comercio
Educación
Transporte
Medio Ambiente
Telecomunicaciones

Los Usos de Datos refieren al propósito para el cual las organizaciones utilizan los datos (p. ej. para dar soporte a la operativa de la organización, para realizar planificación estratégica). La Tabla 6.4 presenta un conjunto inicial de usos de datos a ser considerados en el *framework*. Este conjunto puede ser extendido de acuerdo a nuevos requerimientos.

Tabla 6.4: Usos de Datos

Nombre	**Descripción**
Operativa	La colección de datos es utilizada para dar soporte a la operativa de la organización.
Toma de Decisiones	La colección de datos es utilizada para la toma de decisiones en la organización.
Planificación Estratégica	La colección de datos es utilizada para la planificación estratégica en la organización.
Inteligencia Artificial	La colección de datos es utilizada para técnicas de inteligencia artificial.
Datos Abiertos	La colección de datos es utilizada para la publicación de datos abiertos.

El Ejemplo 21 continua con la descripción del escenario presentado en el Ejemplo 20 utilizando los nuevos conceptos.

Ejemplo 21

Vinculación Organizaciones Clientes:	MSP es responsable de la aplicación Certificado de Nacido Vivo Electrónico DNIC es responsable del Servicio Básico de Información MSP utiliza Servicio Básico de Información
Vinculación Organizaciones Colecciones:	MSP es responsable de colección Certificados de Nacido Vivos (CNV) DNIC es responsable de colección Datos Básicos de Ciudadanos (DBC) MSP utiliza colección DBC
Usos de Datos:	DNIC usa colección CNV para operativa MSP usa colección CNV para operativa MSP usa colección DBC para operativa
Dominios de Aplicación:	MSP opera en dominio Salud DNIC opera en dominio Ciudadanía

6.2.3. Entidades de Negocio y Operaciones sobre Datos

La Figura 6.3 presenta los conceptos del escenario de trabajo relacionados a las entidades de negocio y operaciones sobre los datos.

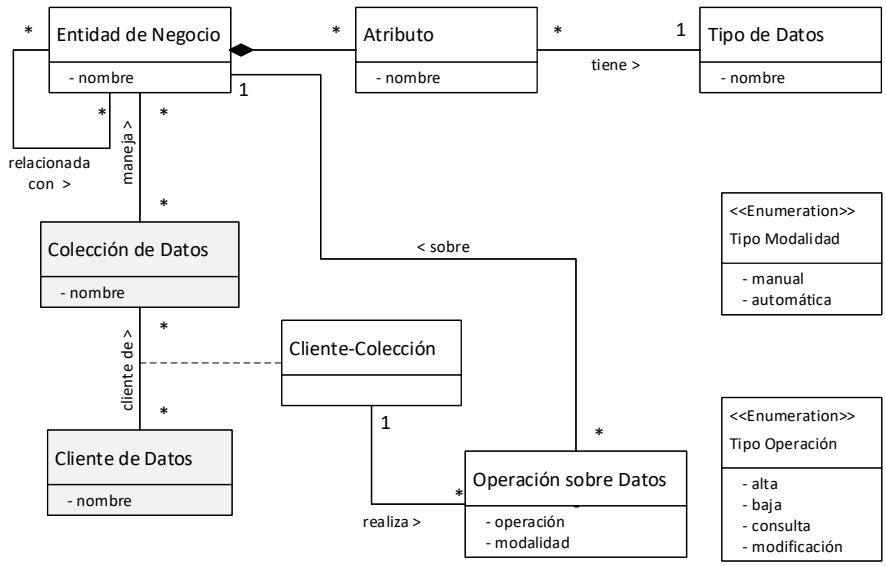


Figura 6.3: Entidades de Negocio y Operaciones sobre Datos

Las Entidades de Negocio representan los elementos conceptuales clave que se manejan en el escenario (p. ej. ciudadano, certificado). Una entidad de negocio puede ser manejada por varias colecciones de datos y estar relacionada con otras entidades de negocio.

Las entidades de negocio (p. ej. ciudadano) tienen atributos (p. ej. nombre, dirección) que pueden ser de distintos tipos (p. ej. alfanumérico, geográfico). La Tabla 6.5 presenta un conjunto inicial de tipos de datos considerados en el *framework*. Este conjunto puede ser extendido de acuerdo a nuevos requerimientos.

Tabla 6.5: Tipos de Datos

Nombre
Alfanumérico
Numérico
Fecha
Geográfico
Imagen

Por otro lado, los clientes de datos pueden realizar distintos tipos de operaciones sobre las entidades manejadas en una colección de datos. En particular, las operaciones que considera el *framework* son alta, baja, consulta y modificación en modalidad tanto manual (i.e. con intervención de una persona) como automática.

El Ejemplo 22 continua con la descripción del escenario presentado en el Ejemplo 20 utilizando los nuevos conceptos.

Ejemplo 22

Entidades:	Certificado de Nacido Vivo (entCNV) Ciudadano
Entidades en Colecciones:	La entidad entCNV se maneja en la colección CNV La entidad Ciudadano se maneja en la colección DBC
Operaciones sobre Datos:	La aplicación appCNVE realiza todas las operaciones sobre la entidad entCNV en la colección CNV La aplicación appCNVE realiza operaciones de consulta sobre la entidad Ciudadano en la colección DBC

6.2.4. Actores de Datos

La Figura 6.4 presenta los conceptos del escenario de trabajo relacionados a actores de datos[2].

Los Actores de Datos refieren a actores clave en el marco del escenario (p. ej. personas, grupos de personas). Estos actores pueden estar vinculados a organizaciones, relacionándose de forma similar que éstas con las colecciones y clientes de datos.

[2] Algunas relaciones entre conceptos no se incluyen para simplificar el diagrama.

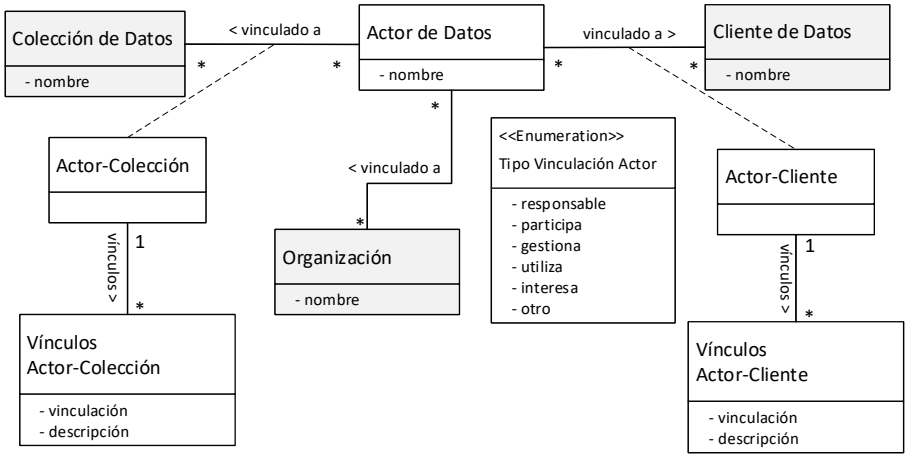


Figura 6.4: Actores de Datos

El Ejemplo 23 presenta posibles actores de datos en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 23

Actores de Datos:	Salud.uy Gerentes MSP Médicos de Proveedor A
Vinculación Actores Colecciones:	Salud.uy está interesado en colección Certificados de Nacido Vivo (CNV) Gerentes MSP están interesados en colección CNV Médicos de Proveedor A utilizan colección CNV
Vinculación Actores Clientes:	Médicos de Proveedor A utilizan aplicación appCNVE Gerentes MSP utilizan aplicación appCNVE

6.2.5. Procesos de Negocio

La Figura 6.5 presenta los conceptos del escenario de trabajo relacionados a procesos de negocio[3].

Los Procesos de Negocio refieren a procesos de alto nivel que son llevados a cabo por los distintos actores para abordar las necesidades de negocio del escenario. Estos procesos están muchas veces implícitos en el escenario o se describen de manera informal (p. ej. textual).

[3] Algunas relaciones entre conceptos no se incluyen para simplificar el diagrama.

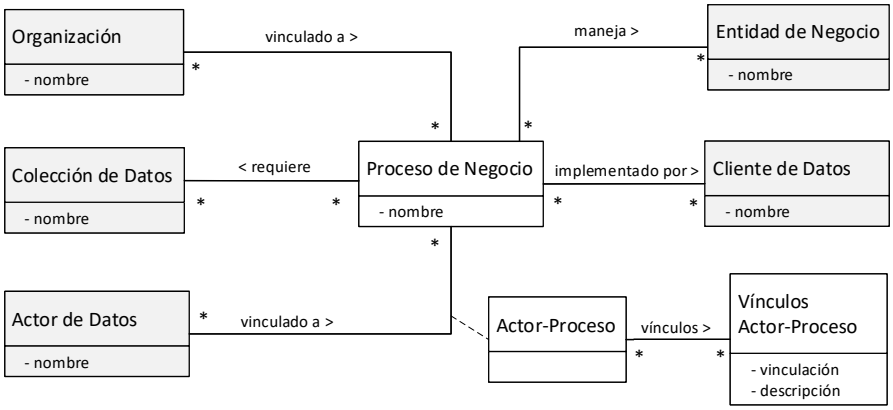


Figura 6.5: Procesos de Negocio

La implementación de los procesos de negocio está dada por los distintos clientes de datos (p. ej. por un proceso implementado en un BPMS[4], por un conjunto de aplicaciones). Además, los procesos de negocio manejan entidades de negocio, requieren colecciones de datos y se vinculan con organizaciones así como con actores de datos.

El Ejemplo 24 presenta el proceso de negocio clave en el marco de la generación de certificados de nacido vivo en Uruguay.

Ejemplo 24

Procesos de Negocio:	Certificado de Nacido Vivo (procCNV)
Clientes para Implementación:	appCNVE implementa procCNV servBI implementa procCNVE
Colecciones Requeridas:	Base de Datos CNV Flujo de Datos DBC
Entidades de Negocio:	Certificado de Nacido Vivo Ciudadano
Vínculos con Organizaciones:	DNIC vinculado a procCNV MSP vinculado a procCNV
Vínculos con Actores:	Salud.uy interesado en procCNV Gerentes MSP interesados en procCNV

[4] Business Process Management System

6.3. Actividades de la Etapa

Esta sección detalla las actividades de la caracterización técnica y de negocio del escenario, las cuales se presentan gráficamente en la Figura 6.6.

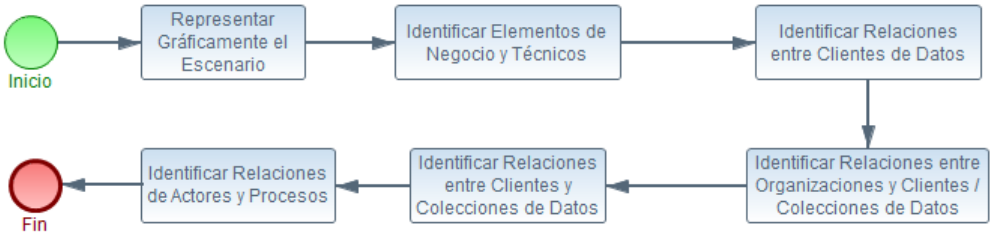


Figura 6.6: Actividades de la Caracterización Técnica y de Negocio

6.3.1. Representar Gráficamente el Escenario

La primera actividad de la etapa de caracterización técnica y de negocio del escenario consiste en representar gráficamente el mismo. Esta representación brinda una visión general del escenario y facilita identificar los distintos elementos de la caracterización técnica y de negocio.

Para la representación gráfica del escenario se sugiere utilizar los elementos gráficos de la Figura 1.2, los cuales se utilizan también para describir el escenario del caso de estudio en la Figura 4.3 y la Figura 4.4.

A modo de ejemplo, la Figura 6.7 presenta la representación gráfica del escenario de trabajo utilizado en los ejemplos de la Sección 6.2, el cual se enmarca en el proceso de negocio que permite la generación de certificados de nacido vivo en Uruguay.

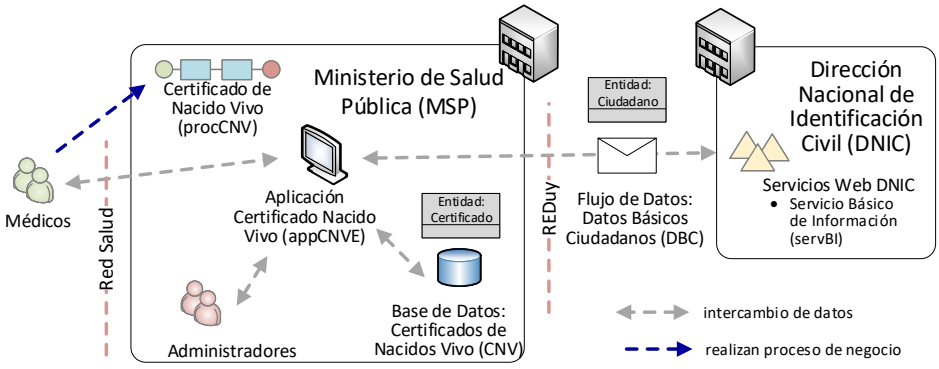


Figura 6.7: Representación Gráfica del Escenario

6.3.2. Identificar Elementos de Negocio y Técnicos del Escenario

La segunda actividad de la etapa de caracterización técnica y de negocio del escenario consiste identificar los elementos de negocio y técnicos del mismo, en base a la representación gráfica elaborada en la actividad anterior. En particular, estos elementos de negocio incluyen:

- las organizaciones que participan en el escenario
- los dominios de aplicación en los que operan estas organizaciones (p. ej. salud, energía)
- las colecciones de datos de interés
- los clientes de datos de interés, en particular, los procesos de negocio
- las entidades y procesos de negocio de interés
- los actores de datos

Se sugiere registrar los elementos identificados en una planilla con la estructura que se presenta en la Tabla 6.6. Para los elementos que lo requieran se sugiere además elaborar planillas complementarias con información más detallada.

Tabla 6.6: Elementos de Negocio y Técnicos del Escenario de Trabajo.

Conceptos	**Elementos Identificados**
Organizaciones	Organización 1, Organización 2
Dominios de Aplicación	Dominio 1, Dominio 2
Colección de Datos	Colección de Datos 1, Colección de Datos 2, Colección de Datos 3
Clientes de Datos	
Servicios	S1, S2, S3
Procesos	Proceso 1, Proceso 2
Aplicaciones	Aplicación 1
Sistemas	Sistema 1
Roles de Usuario	Rol de Usuario 1, Rol de Usuario 2
Entidades de Negocio	Entidad 1, Entidad 2
Procesos de Negocio	Proceso de Negocio 1, Proceso de Negocio 2
Actores de Datos	Actor de Datos 1, Actor de Datos 2

El Ejemplo 25 presenta la utilización de esta planilla para el registro de los elementos de negocio y técnicos identificados en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 25

Conceptos	Elementos Identificados
Organizaciones	MSP, DNIC
Dominios de Aplicación	Salud, Ciudadanía
Colección de Datos	Base de Datos CNV, Flujo de Datos DBC
Clientes de Datos	
Servicios	Servicio Básico de Información (servBI)
Procesos	
Aplicaciones	Aplicación Certificado Nacido Vivo Electrónico (appCNVE)
Sistemas	
Roles de Usuario	Administrador, Médico
Entidades de Negocio	Certificado, Ciudadano
Procesos de Negocio	Proceso de Certificado de Nacido Vivo (procCNV)
Actores de Datos	Salud.uy, Gerentes MSP, Médicos de Proveedor A

6.3.3. Identificar Relaciones entre Clientes de Datos

La tercera actividad de la etapa de caracterización técnica y de negocio del escenario consiste en identificar las relaciones entre los clientes de datos (p. ej. servicios, aplicaciones). Estas relaciones son de suma importancia para conocer la traza de los datos. La Figura 6.8 muestra la parte del modelo conceptual que representa estas relaciones.

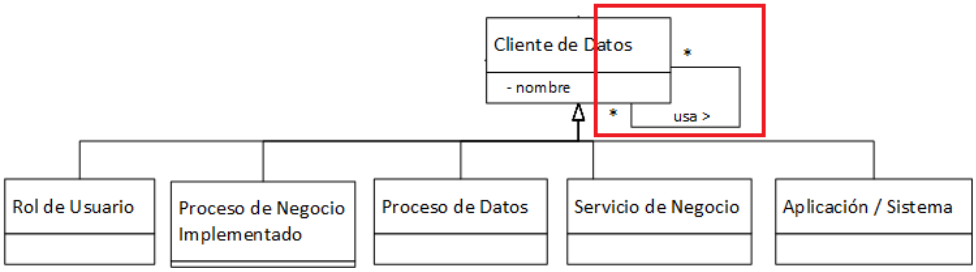


Figura 6.8: Relaciones Entre Clientes de Datos

Se sugiere registrar las relaciones entre clientes de datos identificadas en una planilla con la estructura que se presenta en la Tabla 6.7, donde se especifica qué clientes de datos usa cada cliente de la primera columna.

Tabla 6.7: Relaciones entre Clientes de Datos.

	S1	S2	Proceso 1	Proceso 2	Aplicación 1	Sistema 1	Rol 1	Rol 2
S1		X						
S2								
Proceso 1		X		X		X		
Proceso 2	X							
Aplicación 1	X							
Sistema 1	X	X						
Rol 1					X	X		
Rol 2					X	X		

El Ejemplo 26 presenta la utilización de esta planilla para el registro de las relaciones entre clientes de datos identificadas en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 26

	servBI	appCNVE	Administrador	Médico
servBI				
appCNVE	X			
Administrador		X		
Médico		X		

6.3.4. Relaciones entre Organizaciones y Clientes / Colecciones de Datos

La cuarta actividad de la etapa de caracterización técnica y de negocio del escenario consiste en identificar y describir cómo se vinculan las organizaciones con los clientes y colecciones de datos. La Figura 6.9 muestra la parte del modelo conceptual correspondiente a estas relaciones.

Se sugiere registrar la información de las relaciones entre organizaciones y clientes de datos en una planilla con la estructura que se presenta en la Tabla 6.8.

Tabla 6.8: Relaciones entre Clientes de Datos y Organizaciones

Clientes de Datos	Organizaciones	
	Organización 1	**Organización 2**
Cliente de Datos 1	responsable (descripción)	participa (descripción)
Cliente de Datos 2	participa (descripción)	responsable (descripción)
Cliente de Datos 3	responsable (descripción)	participa (descripción)
...		

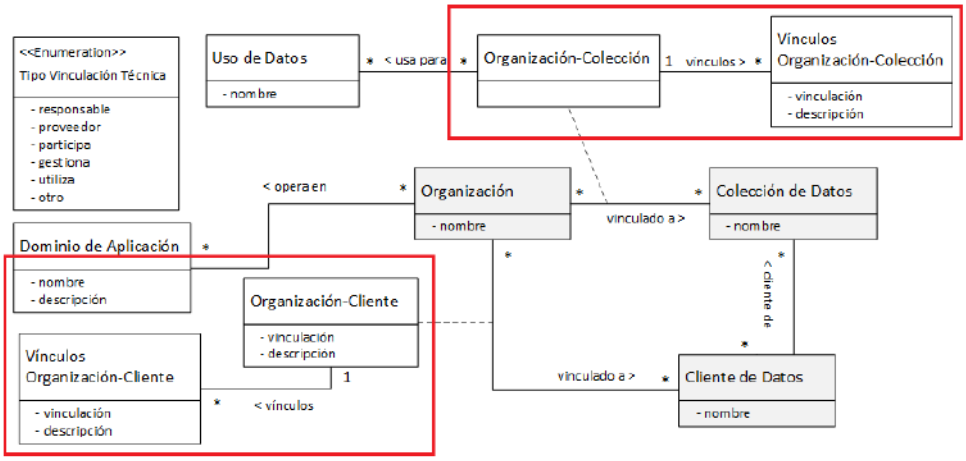


Figura 6.9: Relaciones entre Organizaciones y Clientes / Colecciones de Datos

El Ejemplo 27 presenta la utilización de esta planilla para el registro de las relaciones entre organizaciones y clientes de datos identificadas en el marco del proceso de negocio de certificado de nacido vivo en Uruguay[5].

Ejemplo 27

Clientes de Datos	**Organizaciones**	
	MSP	**DNIC**
servBI	utiliza	responsable
appCNVE	responsable, utiliza	
Médico		
Administrador	responsable	

Por otro lado, se sugiere registrar la información de las relaciones entre organizaciones y colecciones de datos en una planilla con la estructura que se presenta en la Tabla 6.9

Tabla 6.9: Relaciones entre Colecciones de Datos y Organizaciones

Colecciones de Datos	**Organizaciones**	
	Organización 1	**Organización 2**
Colección de Datos 1	responsable (descripción)	participa (descripción)
Subcolección de Datos 1.1	responsable (descripción)	participa (descripción)
Colección de Datos 2	participa (descripción)	responsable (descripción)

[5]No se incluye la descripción para simplificar la planilla

El Ejemplo 28 presenta la utilización de esta planilla para el registro de las relaciones entre organizaciones y colecciones de datos identificadas en el marco del proceso de negocio de certificado de nacido vivo en Uruguay[6].

Ejemplo 28

Colecciones de Datos	Organizaciones	
	MSP	**DNIC**
Base de Datos CNV	responsable	participa
Flujo de Datos DBC	utiliza	responsable

6.3.5. Identificar Relaciones entre Clientes y Colecciones de Datos

La quinta actividad de la etapa de caracterización técnica y de negocio del escenario consiste en identificar sobre qué colecciones de datos realizan operaciones los clientes de datos. Además, se debe especificar qué tipos de operaciones son realizadas (p. ej. altas, consultas, modificaciones) sobre las entidades manejadas en dichas colecciones. La Figura 6.10 muestra la parte del modelo conceptual que representa estas relaciones.

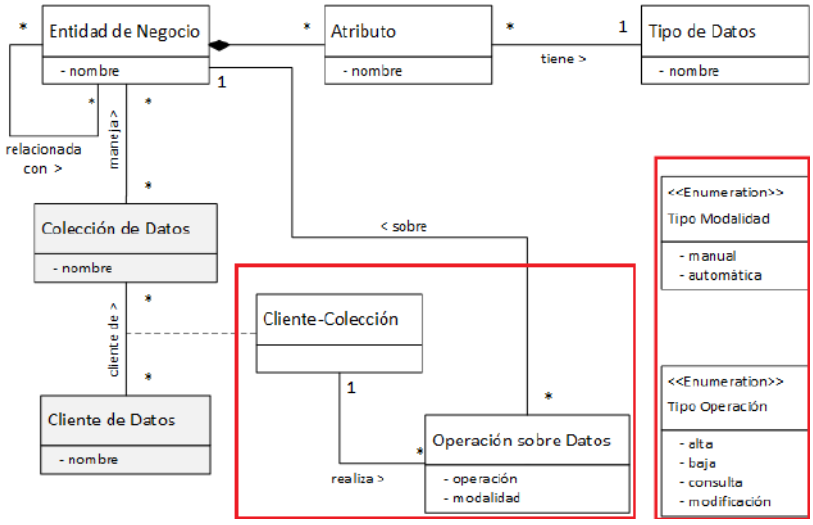


Figura 6.10: Relaciones entre Clientes y Colecciones de Datos

Se sugiere registrar la información de las relaciones entre clientes y colecciones de datos en una planilla con la estructura que se presenta en la Tabla 6.10.

[6]No se incluye la descripción para simplificar la planilla

Tabla 6.10: Relaciones entre Clientes y Colecciones de Datos

Colecciones de Datos	**Clientes de Datos**	
	Cliente de Datos 1	**Cliente de Datos 2**
Colección de Datos 1		
Entidad 1	alta (manual), modificación (manual)	baja (automática)
Entidad 2	consulta (manual)	
Colección de Datos 2		
Entidad 3	alta (manual), modificación (manual)	baja (manual)
...		

El Ejemplo 29 presenta la utilización de esta planilla para el registro de las relaciones entre clientes y colecciones de datos identificadas en el marco del proceso de negocio de certificado de nacido vivo en Uruguay[7].

Ejemplo 29

Colecciones de Datos	**Clientes de Datos**	
	servBI	**appCNVE**
Base de Datos CNV		
Certificado		alta (manual), consulta (manual)
Flujo de Datos DBC		
Ciudadano	alta (manual)	consulta (manual)

6.3.6. Identificar Relaciones de Actores y Procesos

La sexta actividad de la etapa de caracterización técnica y de negocio del escenario consiste en identificar las relaciones de los actores de datos y procesos de negocio con los distintos elementos del escenario.

En particular, se debe identificar cómo se relacionan los actores de datos con las colecciones y clientes de datos. Se sugiere registrar la información de estas relaciones en una planilla con la estructura que se presenta en la Tabla 6.11.

Tabla 6.11: Relaciones entre Actores y Clientes / Colecciones de Datos

Colecciones y Clientes de Datos	**Actores de Datos**	
	Actor 1	**Actor 2**
Colección de Datos 1	responsable	utiliza
Colección de Datos 2	utiliza	
Cliente de Datos 1	responsable	utiliza
Cliente de Datos 2		responsable
...		

[7] No se incluyen los clientes de datos Médico y Administrador para simplificar la planilla

El Ejemplo 30 presenta la utilización de esta planilla para el registro de las relaciones identificadas de los actores de datos con las colecciones y clientes de datos, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 30

Colecciones y Clientes de Datos	**Actores de Datos**		
	Salud.uy	**Gerentes MSP**	**Médicos Proveedor A**
Base de Datos CNV	interesa	utiliza	utiliza
appCNVE	interesa	utiliza	utiliza

Asimismo, también se deben identificar las relaciones de los procesos de negocio con los distintos elementos del escenario. Se sugiere registrar esta información en una planilla con la estructura que se presenta en la Tabla 6.12.

Tabla 6.12: Relaciones de Procesos de Negocio

	Procesos de Negocio	
	Proceso de Negocio 1	**Proceso de Negocio 2**
Clientes de Datos	cliente 1	cliente 2
Colecciones de Datos	Colección 1, Colección 2	Col1
Organizaciones	Organización 1	Organización 2
Actores de Datos	Actor 1	Actor 1, Actor 2
Entidades de Negocio	Entidad 1, Entidad 2	Entidad 3

El Ejemplo 31 presenta la utilización de esta planilla para el registro de las relaciones del proceso de negocio identificado en el marco de la generación de certificados de nacido vivo en Uruguay.

Ejemplo 31

	Procesos de Negocio
	Proceso de Certificado de Nacido Vivo
Clientes de Datos	appCNVE, servBI
Colecciones de Datos	Base de Datos CNV, Flujo de Datos DBC
Organizaciones	DNIC, MSP
Actores de Datos	Salud.uy, Gerentes MSP, Médicos de Proveedor A
Entidades de Negocio	Certificado de Nacido Vivo, Ciudadano

6.4. Aplicación en el Caso de Estudio

Esta sección presenta la caracterización técnica y de negocio del escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4), siguiendo las actividades definidas en la Sección 6.3. En particular, se hace foco únicamente en las funcionalidades de gestión de reclamos en el marco de una de las intendencias (i.e. Intendencia de Montevideo, IM).

6.4.1. Representar Gráficamente el Escenario

En esta actividad se representa gráficamente el escenario de trabajo planteado. En este caso se aprovecha la descripción del escenario presentada en la Figura 4.3 de la Sección 4.2. La Figura 6.11 muestra esta representación gráfica.

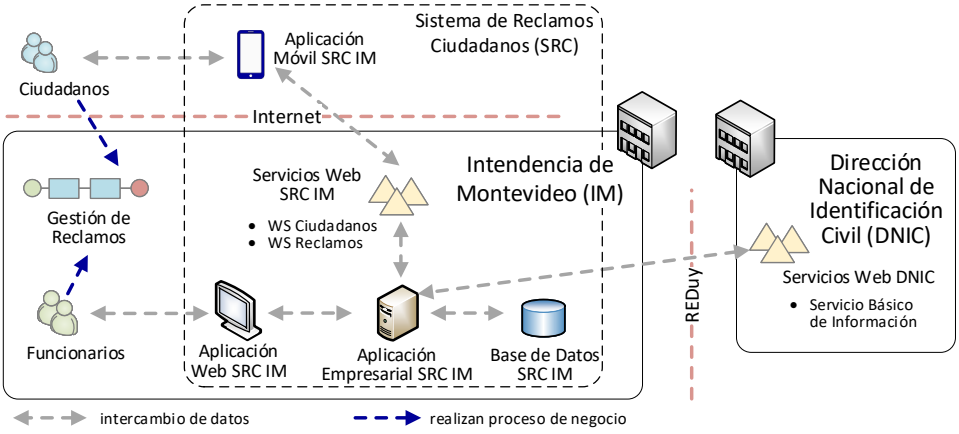


Figura 6.11: Representación Gráfica del Escenario del Caso de Estudio

6.4.2. Identificar Elementos de Negocio y Técnicos del Escenario

En esta actividad se identifican los elementos de negocio y técnicos del escenario, utilizando la representación gráfica elaborada en la actividad anterior. La Tabla 6.13 presenta estos elementos.

Tabla 6.13: Elementos de Negocio y Técnicos del Escenario de Trabajo.

Conceptos	**Elementos Identificados**
Organizaciones	Intendencia de Montevideo (IM), Dirección Nacional de Identificación Civil (DNIC)
Dominio de Aplicación	Ciudadanía, Servicios Públicos
Colección de Datos	Base de Datos SRC IM
Clientes de Datos	
Servicios	WS Ciudadanos IM, WS Reclamos IM, Servicio Básico de Información
Procesos	
Aplicaciones	Aplicación Móvil SRC IM, Aplicación Web SRC IM, Aplicación Empresarial SRC IM
Sistemas	
Roles de Usuarios	Ciudadano, Funcionario IM
Entidades de Negocio	Ciudadano, Reclamo
Procesos de Negocio	Gestión de Reclamos
Actores de Datos	Ciudadanos Zonas Costeras, Funcionarios IM Ventanilla, Funcionarios IM Áreas, Gerentes IM, Sección Evaluación y Monitoreo IM

6.4.3. Identificar Relaciones entre Clientes de Datos

En esta actividad se identifican las relaciones entre clientes de datos. La Tabla 6.14 presenta las relaciones identificadas.

Tabla 6.14: Relaciones entre Clientes de Datos.

	WS Ciudadanos IM	WS Reclamos IM	Servicio Básico de Información	Aplicación Móvil SRC IM	Aplicación Web SRC IM	Aplicación Empresarial SRC IM	Ciudadano	Funcionario IM
WS Ciudadanos IM						X		
WS Reclamos IM						X		
Servicio Básico de Información								
Aplicación Móvil SRC IM	X	X						
Aplicación Web SRC IM						X		
Aplicación Empresarial SRC IM			X					
Ciudadano				X				
Funcionario IM					X			

6.4.4. Relaciones entre Organizaciones y Clientes / Colecciones de Datos

En esta actividad se identifican las relaciones entre organizaciones y clientes / colecciones de datos. La Tabla 6.15 presenta las relaciones entre organizaciones y clientes de datos, mientras que la Tabla 6.16 presenta las relaciones entre organizaciones y colecciones de datos 6.16.

Tabla 6.15: Relaciones entre Clientes de Datos y Organizaciones

Clientes de Datos	**Organizaciones**	
	IM	**DNIC**
WS Ciudadanos IM	responsable	
WS Reclamos IM	responsable	
Servicio Básico de Información	utiliza	responsable
Aplicación Móvil SRC IM	responsable	
Aplicación Web SRC IM	responsable, utiliza	
Aplicación Empresarial SRC IM	responsable, utiliza	
Ciudadano		
Funcionario IM	responsable	

Tabla 6.16: Relaciones entre Organizaciones y Colecciones de Datos

Colecciones de Datos	Organizaciones	
	IM	DNIC
Base de Datos SRC IM	responsable	

6.4.5. Identificar Relaciones entre Clientes y Colecciones de Datos

En esta actividad se identifican las relaciones entre clientes y colecciones de datos. La Tabla 6.17 presenta estas relaciones.

Tabla 6.17: Relaciones entre Clientes y Colecciones de Datos

Clientes de Datos	Colecciones	
	Base de Datos SRC IM	
	Ciudadano	Reclamo
WS Ciudadanos IM		
WS Reclamos IM		
Servicio Básico de Información		
Aplicación Móvil SRC IM		
Aplicación Web SRC IM		
Aplicación Empresarial SRC IM	alta (manual) baja (manual) modificación (manual) consulta (manual)	alta (manual) baja (manual) modificación (manual) consulta (manual)
Ciudadano		
Funcionario IM		

6.4.6. Identificar Relaciones de Actores y Procesos

En esta actividad se identifican las relaciones de actores de datos y procesos de negocio. La Tabla 6.18 presenta las relaciones de actores de datos (G: gestiona, I: interesa, U: utiliza). La Tabla 6.19 presenta las relaciones de los procesos de negocio.

Tabla 6.18: Relaciones de Actores de Datos

Colecciones y Clientes de Datos	**Actores de Datos**				
	Ciudadanos Zonas Costeras	**Funcionarios IM Ventanilla**	**Funcionarios IM Áreas**	**Gerentes IM**	**Sección Evaluación y Monitoreo IM**
Colecciones de Datos					
Base de Datos SRC IM	G	G	G	I	U
Clientes de Datos					
WS Ciudadanos IM					
WS Reclamos IM					
Servicio Básico de Información					
Aplicación Móvil SRC IM	U			I	
Aplicación Web SRC IM		U	U	I	
Aplicación Empresarial SRC IM					
Ciudadano				I	I
Funcionario IM				I	I

Tabla 6.19: Relaciones de Procesos de Negocio

	Procesos de Negocio
	Gestión de Reclamos
Clientes de Datos	Aplicación Web SRC IM Aplicación Empresarial SRC IM WS Ciudadanos IM WS Reclamos IM Servicio Básico de Información Aplicación Móvil SRC IM Funcionario Funcionario IM
Colecciones de Datos	Base de Datos SRC IM
Organizaciones	IM, DNIC
Actores de Datos	Ciudadanos Zonas Costeras (utiliza, interesa) Funcionarios IM Ventanilla (utiliza, interesa) Funcionarios IM Áreas (utiliza, interesa) Gerentes IM (participa, interesa) Sección Evaluación y Monitoreo IM (interesa)
Entidades de Negocio	Reclamo, Ciudadano

7 Caracterización de Calidad de Datos

Continuando con la etapa Caracterizar Escenario del proceso, este capítulo describe los aspectos de calidad de datos de la caracterización.

La Sección 7.1 presenta los objetivos y resultados esperados de la caracterización de calidad. La Sección 7.2 describe el marco conceptual asociado a esta caracterización. La Sección 7.3 detalla las actividades a realizar en la etapa y la Sección 7.4 describe estas actividades en el marco del caso de estudio presentado en el Capítulo 4.

7.1. Objetivos y Resultados Esperados

El objetivo de esta etapa es identificar los elementos relevantes para el escenario de trabajo desde un punto de vista de la calidad de datos. En esta etapa se deben identificar, por ejemplo, los requerimientos y problemas de calidad de datos relevantes para el escenario.

La caracterización de calidad del escenario constituye un insumo importante para etapas posteriores del proceso. En particular, esta caracterización es fundamental para la definición de la estrategia de gestión de calidad de datos así como para la definición del plan de mejora.

Los principales roles del CCD involucrados en esta etapa de caracterización de calidad son el Experto de Negocio y el Analista de Calidad de Datos.

El resultado esperado de la etapa es la identificación de los siguientes elementos del escenario:

- requerimientos de calidad de datos
- relaciones entre requerimientos y elementos del escenario
- problemas de calidad de datos

7.2. Marco Conceptual Asociado

Esta sección describe los principales conceptos asociados a la caracterización de calidad del escenario. Estos conceptos permiten y facilitan describir aspectos de calidad de datos de los escenarios en los que se puede aplicar el *framework*. En particular, estos conceptos permiten especificar requerimientos y problemas de calidad de datos, así como sus relaciones con otros elementos del escenario.

7.2.1. Requerimientos de Calidad de Datos

La Figura 7.1 presenta los conceptos del escenario de trabajo relacionados a requerimientos de calidad de datos[1].

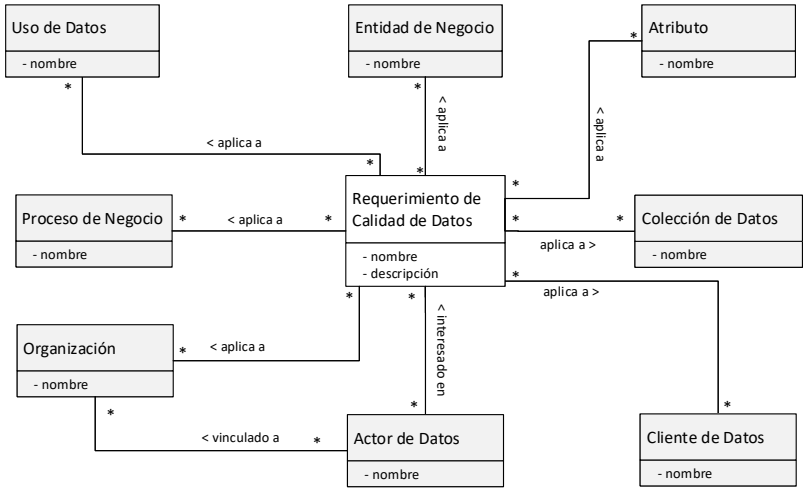


Figura 7.1: Requerimientos de Calidad de Datos

Los Requerimientos de Calidad de Datos refieren a requerimientos que deben cumplir los datos en el marco del escenario y son de interés para distintos actores de datos. Estos requerimientos pueden surgir de leyes, estándares y políticas internas de una organización, entre otros.

Los requerimientos de calidad de datos pueden entonces aplicar globalmente a un escenario así como a elementos específicos del mismo: organizaciones, colecciones de datos, clientes de datos, procesos de negocio, entidades de negocio y atributos de entidades de negocio.

[1] Algunas relaciones entre conceptos no se incluyen para simplificar el diagrama.

El Ejemplo 23 presenta posibles requerimientos de calidad de datos en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 32

Requerimiento de Calidad de Datos: La fecha de nacimiento en un certificado de nacido vivo no puede ser posterior a la actual.

Requerimiento de Calidad de Datos: El número de serie de un certificado de nacido vivo no puede ser nulo.

7.2.2. Problemas de Calidad de Datos

La Figura 7.2 presenta los conceptos del escenario de trabajo relacionados a problemas de calidad de datos.

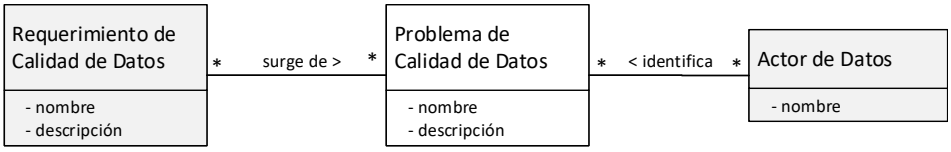


Figura 7.2: Problemas de Calidad de Datos

Los problemas de calidad de datos refieren a cuestiones que se identifican como problemas de los datos (p. ej. las direcciones de correo postal de los clientes no están actualizadas) y que generalmente tienen consecuencias visibles para las organizaciones (p. ej. más del 50 % de los clientes no reciben las facturas). Estos problemas son identificados principalmente por actores de datos.

El Ejemplo 23 presenta posibles problemas de calidad de datos que pueden existir en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 33

Problema de Calidad de Datos: Los usuarios con rol administrador de la aplicación Certificado de Nacido Vivo Electrónico han reportado que un gran número de certificados no tienen datos completos de la madre.

Problema de Calidad de Datos: Se pudo comprobar que los países se especifican de distinta forma en los certificados almacenados en la base de datos. Por ejemplo, Uruguay aparece de las siguientes formas: Uruguay, URU y UY.

7.3. Actividades de la Etapa

Esta sección detalla las actividades de la caracterización de calidad del escenario, las cuales se presentan gráficamente en la Figura 7.3.

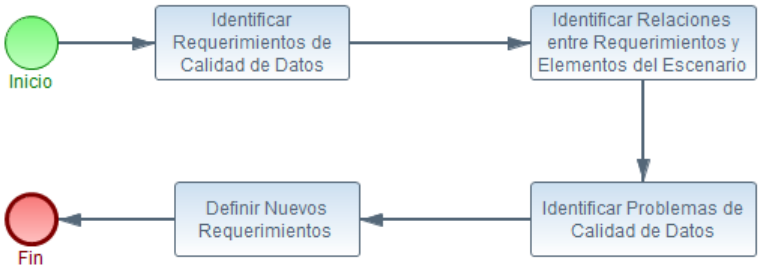


Figura 7.3: Actividades de la Caracterización de Calidad

7.3.1. Identificar Requerimientos de Calidad de Datos

La primera actividad de la etapa de caracterización de calidad del escenario consiste en identificar:

- requerimientos de calidad de datos
- interés de actores en estos requerimientos

Se sugiere registrar los elementos identificados en una planilla con la estructura que se presenta en la Tabla 7.1. Para los elementos que lo requieran se sugiere además elaborar planillas complementarias con información más detallada.

Tabla 7.1: Requerimientos de Calidad de Datos

ID	Requerimiento
ID 1	Requerimiento 1
ID 2	Requerimiento 2.

El Ejemplo 34 presenta la utilización de esta planilla para el registro de los requerimientos de calidad de datos identificados, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 34

ID	Requerimiento
RQ1	Fecha de nacimiento CNV no puede ser posterior a la actual.
RQ2	Número de serie de certificado no puede ser nulo.

Asimismo, en este actividad se deben identificar los actores de datos que tienen interés en los requerimientos identificados. Se sugiere registrar la información de las relaciones identificadas en una planilla con la estructura que se presenta en la Tabla 7.2.

Tabla 7.2: Interés de Actores de Datos en Requerimientos

	Requerimientos	
	Req 1	**Req 2**
Actores de Datos		
Actor 1		X
Actor 2	X	X

El Ejemplo 35 presenta la utilización de esta planilla para el registro del interés de actores en requerimientos, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 35

	Requerimientos	
	RQ1	**RQ2**
Actores de Datos		
Salud.uy	X	X
Gerentes MSP	X	X
Médicos Proveedor A	X	X

7.3.2. Relaciones entre Requerimientos y Elementos del Escenario

La segunda actividad de la etapa de caracterización de calidad consiste en identificar sobre qué elementos del escenario (p. ej. entidades, colecciones de datos, clientes de datos) impactan los requerimientos de calidad de datos identificados.

Se sugiere registrar la información de las relaciones identificadas en una planilla con la estructura que se presenta en la Tabla 7.3.

Tabla 7.3: Relaciones entre Requerimientos de Calidad de Datos y Elementos del Escenario

Elementos	**Requerimientos**	
	Req 1	**Req 2**
Entidades de Negocio		
Entidad 1		X
Entidad 2	X	
Atributos		
Atributo 1		X
Atributo 2	X	X
Colecciones de Datos		
Colección de Datos 1	X	
Colección de Datos 2	X	
Clientes de Datos		
Cliente de Datos 1	X	
Cliente de Datos 2	X	X
Organizaciones		
Organización 1	X	X
Organización 2	X	X
Procesos de Negocio		
Proceso de Negocio 1	X	X
Proceso de Negocio 2	X	X

El Ejemplo 36 presenta la utilización de esta planilla para el registro de las relaciones entre los requerimientos y elementos del escenario, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 36

Elementos	**Requerimientos**	
	RQ1	**RQ2**
Entidades de Negocio		
Certificado	X	X
Ciudadano		
Colecciones de Datos		
Base de Datos CNV	X	X
Flujo de Datos DBC		
Clientes de Datos		
procCNV	X	X
appCNVE	X	X
servBI		
Organizaciones		
MSP	X	X
DNIC		
Procesos de Negocio		
Proceso de Certificado de Nacido Vivo	X	X

7.3.3. Identificar Problemas de Calidad de Datos

La tercera actividad de la etapa de caracterización de calidad consiste en identificar problemas de calidad de datos. Para esto se sugiere realizar entrevistas a los diferentes actores identificados, ya que cada uno tiene una visión diferente de los problemas de calidad dependiendo fuertemente de su contexto de trabajo.

Se sugiere registrar los problemas identificados en una planilla con la estructura que se presenta en la Tabla 7.4.

Tabla 7.4: Problemas de Calidad de Datos

ID	Problemas
P1	Problema 1
P2	Problema 2

El Ejemplo 37 presenta la utilización de esta planilla para el registro de los problemas de calidad de datos identificados, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 37

ID	Problemas
PR1	Un gran número de certificados no tienen datos completos de la madre.

Asimismo, se sugiere registrar la relación entre los actores y problemas de calidad de datos en una planilla con la estructura que se presenta en la Tabla 7.5.

Tabla 7.5: Problemas de Calidad identificados por Actores

Actores	Problemas identificados		
	P1	P2	P3
Actores de Datos			
Actor 1	X	X	X
Actor 2	X		
Actor 3	X	X	X
Actor 4	X		
...			

En particular, se observa que el problema P1 fue reportado por todos los actores, por lo que es un indicio de que es un problema de relevancia.

El Ejemplo 38 presenta la utilización de esta planilla para el registro de las relaciones entre actores y problemas, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 38

	Problemas identificados
	PR1
Actores de Datos	
Salud.uy	
Gerentes MSP	X
Médicos de Proveedor A	

7.3.4. Definir Nuevos Requerimientos de Calidad de Datos

La cuarta actividad de la etapa de caracterización de calidad consiste en la identificación de nuevos requerimientos de calidad de datos a partir de los problemas de calidad identificados.

Si bien en etapas previas se identifican requerimientos de calidad de datos, estos corresponden a los establecidos por el dominio considerado y/o por las organizaciones participantes. A partir de los problemas de calidad identificados, se pueden identificar nuevos requerimientos. Por ejemplo, teniendo en cuenta el problema P1 que fue reportado por todos los actores, puede surgir un nuevo requerimiento de calidad de datos Req 3.

Se sugiere registrar la relación entre los problemas de calidad y los requerimientos que surgen a partir de ellos en una planilla con la estructura que se presenta en la Tabla 7.6.

Tabla 7.6: Relaciones entre Problemas y Requerimientos de Calidad

Problemas	Requerimientos		
	Req 1	**Req 2**	**Req 3**
P1			X
P2			

El Ejemplo 39 y el Ejemplo 40 presentan cómo se puede registrar un nuevo requerimiento que surge de un problema de calidad de datos, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 39

ID	Requerimiento
RQ3	Los nombres, apellidos y cédula de identidad de la madre en un certificado de nacido vivo deben estar especificados.

Ejemplo 40

Problemas	Requerimientos		
	RQ1	**RQ2**	**RQ3**
PR1			X

7.4. Aplicación en el Caso de Estudio

Esta sección presenta la caracterización de calidad del escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 7.3.

7.4.1. Identificar Requerimientos de Calidad de Datos

En esta actividad se identifican requerimientos de calidad de datos relevantes para el escenario de trabajo. La Tabla 7.7 presenta los requerimientos identificados.

Tabla 7.7: Requerimientos de Calidad de Datos

ID	Requerimiento
R1	El número de documento del ciudadano tiene que corresponder a una cédula válida.
R2	El email del ciudadano debe tener un formato de correo electrónico válido.

Por otro lado, la Tabla 7.8 presenta el interés de los actores de datos en los requerimientos identificados.

Tabla 7.8: Interés de Actores en Requerimientos de Calidad de Datos

Actores	Requerimientos	
	R1	R2
Ciudadanos Zonas Costeras		
Funcionarios IM Ventanilla	X	X
Funcionarios IM Áreas		X
Gerentes IM	X	X
Sección Evaluación y Monitoreo IM	X	X

7.4.2. Relaciones entre Requerimientos y Elementos del Escenario

En esta actividad se identifican los elementos del escenario a los cuales aplican los requerimientos resultantes de la actividad anterior. La Tabla 7.9 presenta las relaciones entre requerimientos y elementos del escenario.

Tabla 7.9: Relaciones entre Requerimientos de Calidad de Datos y Elementos del Escenario

Elementos	**Requerimientos**	
	R1	**R2**
Entidades de Negocio		
Ciudadano	X	X
Reclamo		
Atributos		
Ciudadano.nombres		
Ciudadano.apellidos		
Ciudadano.numDoc	X	
Ciudadano.email		X
Reclamo.id		
Reclamo.motivo		
Reclamo.fecha		
Reclamo.coordX		
Reclamo.coordY		
Colecciones de Datos		
Base de Datos SRC IM	X	X
Clientes de Datos		
WS Ciudadanos IM	X	X
WS Reclamos IM		
Servicio Básico de Información	X	
Proceso Gestión de Reclamos	X	X
Aplicación Móvil SRC IM	X	X
Aplicación Web SRC IM	X	X
Aplicación Empresarial SRC IM	X	X
Ciudadano	X	X
Funcionario IM	X	X
Organizaciones		
IM	X	X
DNIC	X	X
Procesos de Negocio		
Gestión de Reclamos	X	X

7.4.3. Identificar Problemas de Calidad de Datos

En esta actividad se identifican los problemas de calidad de datos relevantes para el escenario. En base al listado inicial de aspectos de calidad de datos presentado en la Sección 4.5, se elabora la planilla que se presenta en la Tabla 7.10.

Tabla 7.10: Problemas de Calidad de Datos

ID	Problemas
P1	**Nombres irreales.** Se han detectado nombres de fantasía, apodos o *nicknames* en los campos destinados a los nombres y apellidos del ciudadano.
P2	**Correos electrónicos inexistentes.** A algunos ciudadanos registrados no les llegan los correos electrónicos que envía el sistema porque ingresaron incorrectamente su dirección de correo.
P3	**Domicilios no encontrados.** Algunos de los domicilios ingresados por los ciudadanos no se corresponden con ninguna dirección oficial de la capa de direcciones que maneja el sistema.
P5	**Edades poco confiables.** Algunas edades de los ciudadanos registrados (que se calculan en base a la fecha de nacimiento ingresada) resultan poco confiables, por estar fuera de los rangos en que deberían encontrarse los usuarios de esta aplicación (mayores de 10 años y menores de 100 años).
P5	**Uso abusivo.** Se han constatado varios casos de usuarios que realizan un uso abusivo del sistema, que incluyen: reclamos de incidentes falsos, múltiples reclamos del mismo usuario sobre el mismo incidente, observaciones o fotos de los reclamos con contenido inapropiado o irrelevante.
P6	**Reclamos duplicados.** Una parte importante del trabajo del funcionario que recibe los reclamos es poder identificar los reclamos de distintos ciudadanos que hacen referencia al mismo problema. Si estos reclamos duplicados no son detectados oportunamente, puede suceder que las áreas reciban reclamos que ya fueron resueltos en base a otros reclamos.
P7	**Reclamos rechazados sin aclaraciones.** Algunos funcionarios no completan las observaciones cuando cambian el estado del reclamo. Esto genera disconformidad en algunos ciudadanos, que ven sus reclamos en estado «rechazado» y no conocen el motivo.
P8	**Inconsistencia entre la categoría del reclamo y su ubicación geográfica.** Se ha constado que a veces la categoría del reclamo no es compatible con su ubicación (p. ej. reclamo de la subcategoría «presencia de cianobacterias en playa» es reportado en una ubicación alejada de la playa).
P9	**Inconsistencia entre la categoría del reclamo y su fecha.** En algunos casos, la fecha en la que se ingresa el reclamo no es compatible con la categoría del reclamo (p. ej. se ingresa un reclamo de la subcategoría «falta de limpieza posterior a feria» en una fecha muy posterior a la realización de dicha feria).

Por otro lado, la planilla de la Tabla 7.11 especifica qué actores identificaron cada problema.

Tabla 7.11: Problemas de Calidad identificados por Actores

Actores	**Problemas identificados**								
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**
Actores de Datos									
Ciudadanos Zonas Costeras							X		
Funcionarios IM Ventanilla	X	X		X	X			X	X
Funcionarios IM Áreas			X			X			
Gerentes IM				X		X	X		
Sección Evaluación y Monitoreo IM							X	X	X

7.4.4. Definir Nuevos Requerimientos de Calidad de Datos

En esta actividad se definen nuevos requerimientos de calidad de datos en base a los problemas identificados en la actividad anterior. La Tabla 7.12 presenta la definición de estos nuevos requerimientos.

Tabla 7.12: Requerimientos de Calidad de Datos

ID	**Requerimiento**
R1	El número de documento del ciudadano tiene que corresponder a una cédula válida.
R2	El email del ciudadano debe tener un formato de correo electrónico válido.
R3	Los nombres y apellidos de los ciudadanos deben ser los reales.
R4	El email del ciudadano debe existir.
R5	El domicilio del ciudadano debe existir.
R6	El teléfono del ciudadano debe existir.
R7	La edad del ciudadano que se registra debe estar entre 10 y 110 años.
R8	El reclamo debe corresponder a un hecho real.
R9	El reclamo no debe estar duplicado.
R10	El estado «rechazado» de un reclamo debería tener una aclaración no vacía.
R11	La localización del reclamo debe ser consistente con su categoría.
R12	La categoría del reclamo debe ser consistente con su fecha.

Por otro lado, en la Tabla 7.13 se especifica el interés de los actores en estos nuevos requerimientos.

Por último, en la Tabla 7.14 se especifica de qué problemas surgieron los nuevos requerimientos.

Tabla 7.13: Interés de Actores en Requerimientos de Calidad de Datos

Actores	**Requerimientos**											
	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**
Ciudadanos Zonas Costeras										X		
Funcionarios IM Ventanilla	X	X	X	X			X	X			X	X
Funcionarios IM Áreas		X			X				X			
Gerentes IM	X	X					X		X	X		
Sección Evaluación y Monitoreo IM	X	X								X	X	X

Tabla 7.14: Relaciones entre Problemas y Requerimientos de Calidad

Problemas	**Requerimientos**											
	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R11**
P1			X									
P2				X								
P3					X							
P4							X					
P5								X				
P6									X			
P7										X		
P8											X	
P9												X

8 Examinar Datos Objetivo

Este capítulo describe la segunda etapa del proceso de gestión de calidad propuesto, la cual consiste en examinar los datos de las colecciones identificadas en la caracterización del escenario. Esta etapa tiene como fin conocer las características de los datos.

La Sección 8.1 presenta los objetivos y resultados esperados de esta etapa y la Sección 8.2 describe el marco conceptual asociado a la misma. La Sección 8.3 detalla las actividades a realizar en la etapa y la Sección 8.4 describe estas actividades en el marco del caso de estudio presentado en el Capítulo 4.

8.1. Objetivos y Resultados Esperados

El objetivo de esta etapa es conocer las características de los datos y realizar una primera estimación de su calidad, así como detectar problemas de calidad de datos y definir nuevos requerimientos a partir de estos.

Una de las principales herramientas de soporte en esta etapa es la aplicación de técnicas de *Data Profiling* [Abe15], las cuales permiten obtener información diversa de los datos.

Los principales roles del CCD involucrados en la etapa Examinar Datos Objetivo son el Experto Técnico, el Experto de Negocio y el Técnico de Calidad de Datos.

El resultado esperado de esta etapa es una descripción de las características principales de los datos objetivo y una primera estimación de su calidad. En particular, se apunta a conocer y/o registrar:

- los tipos de cada colección de datos
- los tipos de cada atributo de las entidades
- una primera estimación del alcance de las problemas identificados
- una primera estimación del cumplimiento de los requerimientos identificados
- nuevos problemas de calidad de datos que se detecten en base a la aplicación de técnicas de *Data Profiling*
- la definición de nuevos requerimientos de calidad de datos que surjan a partir de estos nuevos problemas

8.2. Marco Conceptual Asociado

Esta sección describe los principales conceptos asociados a la etapa Examinar Datos Objetivo. Estos conceptos permiten especificar las características de los datos, los resultados de la aplicación de técnicas de *Data Profiling*, las estimaciones de la calidad de los datos en base a estos resultados, así como problemas que se detecten a partir de los mismos.

La Figura 8.1 presenta los conceptos asociados a estos elementos.

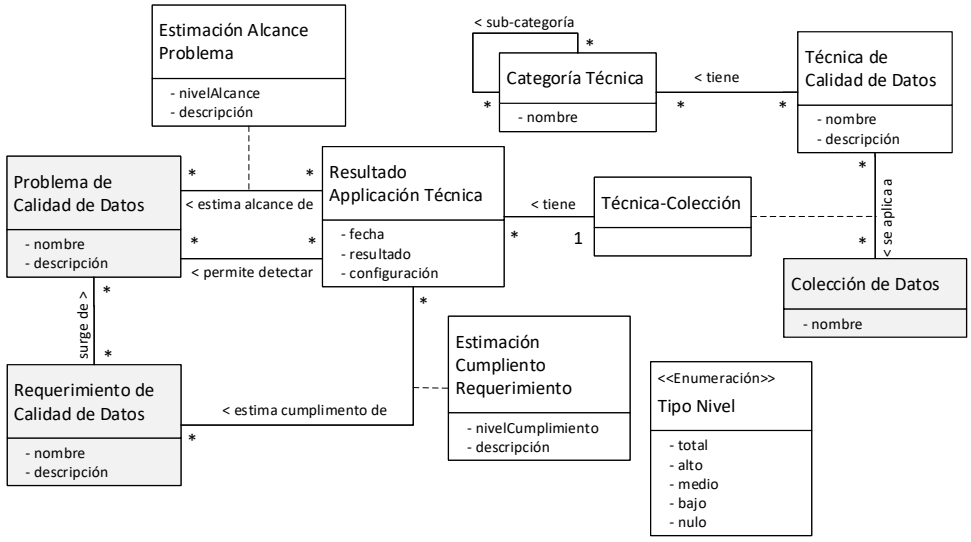


Figura 8.1: Técnicas de Calidad de Datos, Resultados y Estimaciones

Las Técnicas de Calidad de Datos refieren a mecanismos que procesan datos y que están orientados a asistir en un proceso de gestión de calidad de datos. Estos mecanismos pueden tener asociadas distintas categorías de acuerdo a su fin (p. ej. análisis, *Data Profiling*, transformación). En particular, la categoría relevante para esta etapa es la de *Data Profiling*.

Las técnicas de *Data Profiling* se aplican sobre colecciones de datos utilizando una configuración específica (p. ej. estableciendo sobre qué atributos de una entidad se aplican). La Tabla 8.1 presenta algunas técnicas comunes de *Data Profiling* [Abe15].

Tabla 8.1: Técnicas de *Data Profiling*

Nombre Técnica
observación directa
obtener cardinalidades
detectar patrones
obtener distribución de valores
detectar correlaciones
chequear unicidad
detectar valores fuera de rango

A modo de ejemplo, a partir de la aplicación de técnicas de *Data Profiling* sobre las colecciones es posible obtener información sobre:

- el volumen de los datos
- patrones en los datos
- cuántos valores repetidos tiene un dato
- la frecuencia con la que un dato tiene asignado un valor nulo
- la distribución que tiene un dato para los diferentes valores que éste toma
- cuáles son los valores mínimo, medio y máximo que presenta un dato

La aplicación de técnicas de *Data Profiling* sobre colecciones de datos producen resultados. Estos resultados pueden utilizarse para estimar el alcance de los problemas de calidad de datos, así como para identificar nuevos problemas. Estos nuevos problemas pueden a su vez generar requerimientos de calidad de datos adicionales. Los resultados de la aplicación de técnicas de *Data Profiling* pueden utilizarse también para tener una primera estimación del grado de cumplimiento de los requerimientos de calidad identificados.

El Ejemplo 41 presenta cómo el resultado de la aplicación de una técnica de *Data Profiling* puede servir para detectar un problema de calidad de datos, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay. A su vez, este problema de calidad de datos origina la definición de un nuevo requerimiento de calidad de datos.

Ejemplo 41

Técnica de Calidad de Datos: Detectar Patrones (Categoría: *Data Profiling*).

Resultado de Aplicación Técnica: Al aplicar la técnica de detectar patrones a la Base de Datos CNV, se puede comprobar que existen distintas formas de especificar los países: nombre completo, tres letras, dos letras. A modo de ejemplo, Uruguay se especifica como Uruguay, URY y UY.

Problema de Calidad de Datos Detectado: No existe una única forma de especificar los países en los certificados de nacido vivo.

Requerimiento de Calidad de Datos: Los países deben especificarse utilizando los códigos Alpha 3 del estándar ISO 3166-1.

8.3. Actividades de la Etapa

Esta sección detalla las actividades de la etapa Examinar Datos Objetivo, las cuales se presentan gráficamente en la Figura 8.2.

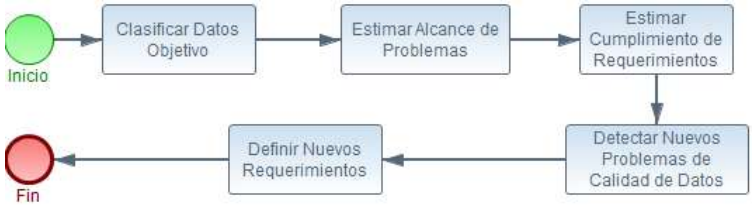


Figura 8.2: Actividades de la Caracterización Técnica

8.3.1. Clasificar Datos Objetivo

La primera actividad de la etapa Examinar Datos Objetivo consiste en clasificar los datos en base su tipo. En particular, se debe especificar el tipo de cada colección de datos (p. ej. base de datos relacional, base de datos documental) y el tipo de cada atributo (p. ej. alfanumérico, imagen, geográfico) identificado en la caracterización del escenario.

La Figura 8.3 y la Figura 8.4 presentan las partes del modelo conceptual correspondiente a estos elementos.

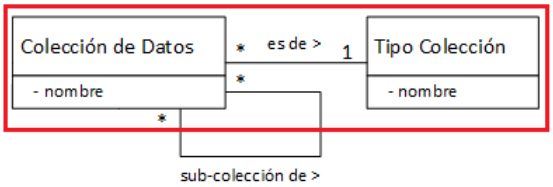


Figura 8.3: Tipos de Colecciones de Datos

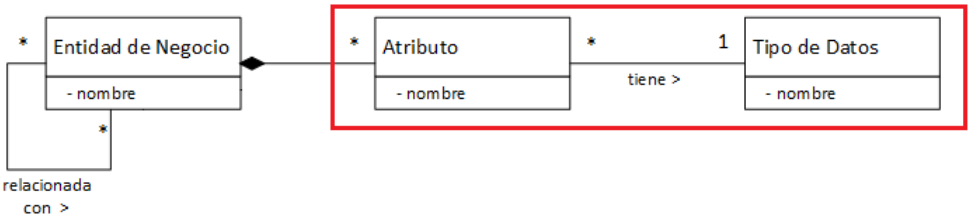


Figura 8.4: Tipos de Datos de los Atributos de Entidades

Se sugiere registrar la información de los tipos de colección en una planilla con la estructura que se presenta en la Tabla 8.2.

Tabla 8.2: Tipos de Colecciones de Datos

Colección de Datos	Tipo de Colección
Colección de Datos 1	BD Relacional
Colección de Datos 2	BD Documental
Colección de Datos 3	Flujo de Datos

Se sugiere registrar la información de los tipos de datos de los atributos en una planilla con la estructura que se presenta en la Tabla 8.3

Tabla 8.3: Tipos de Datos de los Atributos de Entidades

Atributo	Tipo de Datos
Entidad1.Atributo1	Alfanumérico
Entidad1.Atributo2	Numérico
Entidad2.Atributo3	Fecha
Entidad2.Atributo3	Imagen

El Ejemplo 42 y el Ejemplo 43 presentan la utilización de estas planillas en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 42

Colección de Datos	Tipo de Colección
Base de Datos CNV	BD Relacional
Flujo de Datos DBC	Flujo de Datos

Ejemplo 43

Atributo	Tipo de Datos
Certificado.Numero	Alfanumérico
Certificado.Dirección	Geográfico
Certificado.CI_Madre	Alfanumérico
Ciudadano.Nombre	Alfanumérico
Ciudadano.Foto	Imagen

8.3.2. Estimar Alcance de Problemas

La segunda actividad de la etapa Examinar Datos Objetivo consiste en estimar el alcance de cada uno de los problemas identificados en la etapa de caracterización, en base a la aplicación de técnicas de *Data Profiling*.

Se sugiere registrar la información de los resultados de la aplicación de estas técnicas en una planilla con la estructura que se presenta en la Tabla 8.4.

Tabla 8.4: Resultados de Aplicación de Técnicas sobre Colección

Resultados Técnica	
Técnica	Técnica 1
Colección de Datos	Colección de Datos 1
Fecha	Fecha 1
Configuración	**Resultado**
Configuración 1	Resultado 1
Configuración 2	Resultado 2
Configuración 3	Resultado 3

Se sugiere registrar la información de la estimación del alcance de los problemas en una planilla con la estructura que se presenta en la Tabla 8.5.

Tabla 8.5: Estimación de Alcance de Problemas de Calidad de Datos

Resultados Técnicas	**Problemas**		
	Problema 1	**Problema 2**	**Problema 3**
Resultado 1 - Técnica 1	alto	medio	
Resultado 2 - Técnica 1		total	nulo

Para cada escenario se deberá establecer el significado de los distintos alcances que define el *framework* (i.e. alto, medio). Asimismo, cabe notar que pueden existir casos en los cuales no sea posible realizar esta estimación de alcance (p. ej. porque puede resultar muy complejo para esta etapa).

El Ejemplo 44 y el Ejemplo 45 presentan la utilización de estas planillas en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 44

Resultados Técnica	
Técnica	obtener distribución de valores
Colección de Datos	Base de Datos CNV
Fecha	08/08/2019
Configuración	**Resultado**
Atributo País	Uruguay (70), URY (98), UY (22)

Ejemplo 45

Resultados Técnicas	**Problemas**	
	Países especificados de distinta forma	**Datos incompletos**
observación directa 03/10/19	alto	nulo
distribución valores 04/10/19	alto	

8.3.3. Estimar Cumplimiento de Requerimientos

La tercera actividad de la etapa Examinar Datos Objetivo consiste en estimar el grado de cumplimiento de los requerimientos de calidad identificados en la etapa de caracterización, aplicando técnicas de *Data Profiling*.

Se sugiere registrar la información de la aplicación de estas técnicas en planillas con la estructura que se presenta en la Tabla 8.4 y la Tabla 8.6.

Tabla 8.6: Estimación de Cumplimiento de Requerimientos con Técnicas de *Data Profiling*.

Resultados Técnicas	Requerimientos		
	Requerimiento 1	**Requerimiento 2**	**Requerimiento 3**
Resultado Técnica 1	alto		
Resultado Técnica 2		medio	
Resultado Técnica 3			bajo

El Ejemplo 46 presenta la utilización de la planilla de la Tabla 8.6 en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 46

Resultados Técnicas	Requerimientos
	Países en formato Alpha 3 de estándar ISO 3166-1
distribución valores 04/10/19	medio

8.3.4. Detectar Nuevos Problemas de Calidad de Datos

La cuarta actividad de la etapa Examinar Datos Objetivo consiste en detectar nuevos problemas de calidad de datos en las entidades identificadas en la etapa de caracterización, aplicando técnicas de *Data Profiling*.

Si bien esta actividad se realiza en general mediante una exploración libre de los datos, se sugiere que esté guiada por las entidades y atributos identificados en la etapa de caracterización de la siguiente manera:

1. aplicar técnicas de *Data Profiling* para cada atributo de cada entidad
2. aplicar técnicas de *Data Profiling* para pares de atributos de una misma entidad que puedan tener alguna relación
3. aplicar técnicas de *Data Profiling* para pares de entidades que puedan tener alguna relación

Se sugiere registrar la información de la aplicación de estas técnicas en planillas con la estructura que se presenta en la Tabla 8.4 y la Tabla 8.7.

Tabla 8.7: Problemas Detectados con Técnicas de *Data Profiling*.

Resultados Técnicas	Problemas		
	Problema 1	**Problema 2**	**Problema 3**
Resultado Técnica 1	X		
Resultado Técnica 2		X	

El Ejemplo 47 presenta la utilización de la planilla de la Tabla 8.7 en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 47

Resultados Técnicas	Problemas
	Datos incompletos
observación directa 03/10/19	X

Se deben también incluir los nuevos problemas de calidad de datos en la planilla presentada en la Tabla 7.5.

8.3.5. Definir Nuevos Requerimientos

La quinta actividad de la etapa Examinar Datos Objetivo consiste en la definición de nuevos requerimientos de calidad de datos a partir de los problemas de calidad detectados.

De forma similar a lo presentado en la caracterización del escenario, se sugiere registrar esta información en planillas que especifican de qué problemas surgen los requerimientos (cf. Sección 7.3.4).

El Ejemplo 48 presenta un ejemplo de nuevo requerimiento que pueden surgir a partir de un problema detectado utilizando técnicas de *Data Profiling*, en el marco del proceso de negocio de certificado de nacido vivo en Uruguay.

Ejemplo 48

Problemas	Requerimientos
	90 % de completitud
Datos incompletos	X

Se debe también incluir los nuevos requerimientos de calidad de datos en la planilla presentada en la Tabla 7.6.

8.4. Aplicación en el Caso de Estudio

Esta sección presenta la etapa Examinar Datos Objetivos en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4), siguiendo las actividades definidas en la Sección 8.3.

8.4.1. Clasificar Datos Objetivo

En esta actividad se clasifican los datos relevantes para el escenario, especificando los tipo de las colecciones y el tipo de datos de los atributos de las entidades de negocio.

En la Tabla 8.8 se detalla el tipo de la colección de datos «Base de Datos SRC IM» y en la Tabla 8.9 se muestran los tipos de datos de los atributos más relevantes para los requerimientos definidos para el escenario.

Tabla 8.8: Tipos de las Colecciones de Datos

Colección de Datos	Tipo de Colección
Base de Datos SRC IM	BD Relacional

Tabla 8.9: Tipos de Dato de los Atributos de Ciudadano y Reclamo

Atributo	Tipo de Datos
Ciudadano.nombres	Alfanumérico
Ciudadano.apellidos	Alfanumérico
Ciudadano.numDoc	Alfanumérico
Ciudadano.email	Alfanumérico
Reclamo.id	Numérico
Reclamo.motivo	Alfanumérico
Reclamo.fecha	Fecha
Reclamo.coordX	Numérico
Reclamo.coordY	Numérico

8.4.2. Estimar Alcance de Problemas

En esta actividad se estima el alcance de los problemas de calidad identificados, aplicando técnicas de *Data Profiling*. En particular, se realiza una exploración directa de los datos de la colección «Base de Datos SRC IM» para estimar el alcance de los problemas que refieren a los nombres (i.e. P1), email (i.e. P2) y fecha de nacimiento de los ciudadanos (i.e. P3). La Figura 8.5 presenta algunos de estos datos.

ID	nombres	apellidos	numDoc	eMail	telefono	sexo	fechaNac
2	Juan	Perez	3.457.2158	jperez@gmal.com	91234543	M	03/07/1993
3	Felipe	Rodriguez	4.254.7516	ferrodri@homail.com	95457812	M	12/05/1905
1	García		5.384.2634	fvgarcia@montevideo.com.uy	93265498	F	10/08/2001
5	Ana González		1.234.5678	anita@adinet.com.uy	94569218	F	26/09/1975
10		Alejandra	5.945.6546	alej0234@adine.com.uy	93234658		02/08/2010
8		Jorge	4.569.0731	jorgito_25@gmail.com	99875093	M	26/02/2000
7	elcarbone		31684564	carboen@hotmal.com	98769054		25/04/2019
12		vidatrico	45439802	trico987@adinet.com	93456239	F	15/5/1903

Figura 8.5: Datos de Ciudadanos

Esta exploración permite estimar el alcance de los problemas antes mencionados:

- P1: Hay valores nulos y valores extraños en los campos de nombre y apellido de los ciudadanos (p. ej. elcarbone, vidatrico)
- P2: Hay valores en el atributo mail que aunque son casillas bien formadas, tienen errores en los dominios (p. ej. @hotmal.com, @gmal.com)
- P4: Hay valores extraños en las fechas de nacimiento (i.e. años 1903, 1906 y 2019)

La Tabla 8.10 y la Tabla 8.11 presentan cómo se registran estos resultados.

Tabla 8.10: Resultados de Exploración Directa sobre Colección Base de Datos SRC IM

Resultados Técnica	
Técnica	Exploración Directa
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Tabla ciudadanos, atributo nombres, apellidos	elcarbone, vidatrico
2. Tabla ciudadanos, atributo email	@hotmal.com, @gmal.com
3. Tabla ciudadanos, atributo fechaNac	1903, 1906, 2019

Tabla 8.11: Aplicación *Data Profiling* para Problemas de Calidad de Datos

Resultados Técnicas	**Problemas**		
	P1 (nombres)	**P2 (email)**	**P4 (fechaNac)**
Exploración Directa (1)	alto		
Exploración Directa (2)		alto	
Exploración Directa (3)			alto

8.4.3. Estimar Cumplimiento de Requerimientos

En esta actividad se estima el cumplimiento de los requerimientos de calidad identificados, aplicando técnicas de *Data Profiling*.

En este caso se utiliza la técnica de *Pattern Finder* provista por la herramienta Data Cleaner, para verificar el grado de cumplimiento del requerimiento R2 del caso de estudio (i.e. el email del ciudadano debe tener un formato válido). La Figura 8.6 presenta el resultado de aplicar esta técnica sobre el atributo email de la tabla ciudadanos.

Figura 8.6: Resultado de Aplicación de Técnica *Pattern Finder*

La aplicación de esta técnica permite comprobar que el grado de cumplimiento de este requerimiento es total.

La Tabla 8.13 y la Tabla 8.12 presentan cómo se registran estos resultados.

Tabla 8.12: Resultados de *Pattern Finder* sobre Colección Base de Datos SRC IM

Técnica	*Pattern Finder*
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Tabla ciudadanos, atributo email	aaaaaaaaaa@aaaaaaaa.aaa

Tabla 8.13: Aplicar Técnicas *Data Profiling* para Requerimientos de Calidad de Datos

Resultados Técnicas	**Requerimientos**		
	R1 (documento)	**R2 (email)**	**.....**
***Pattern Finder* (1)**		total	

8.4.4. Detectar Nuevos Problemas de Calidad de Datos

En esta actividad se busca detectar nuevos problemas de calidad de datos utilizando técnicas de *Data Profiling*. En particular, para el caso de estudio se utiliza las técnicas provistas por la herramienta DataCleaner y se hace foco en los datos de los reclamos.

En la Figura 8.7 se ve la estructura de datos y algunos datos de ejemplo de los reclamos.

ID_RECLAMO	MOTIVO	FECHA_INGR	ESTADO	FECHA_DESDE	AREA	CATEGORIA	SUBCATEGORIA	COORD_X	COORD_Y
2816	Valentín Alvarez entr	30/12/2016	Finalizado	30/12/2016	Calles y veredas	Viales	Bache	570526.709	6147877.70
2818	Basural fuera conten	1/1/2018	Ingresado	1/1/2018	Limpieza	Estado de los co	Contenedor roto	579076.700	6142999.12
2819	A las viviendas	1/1/2019	Finalizado	22/1/2019	Limpieza	Estado de los co	Solicitar traslado c	578941.188	6137947.96
138019	Cámara principal obs	2/1/2019	Finalizado	2/1/2019	Saneamiento	Conexiones y C	Conexion Obstruid	568710.261	6140716.53
139717	Solicita que se conc	2/1/2017	Finalizado	3/1/2017	Limpieza	Contenedores	No pasó Camión L	579469.879	6141592.64
139718	varios focos	2/1/2018	Finalizado	5/1/2018	Alumbrado	Alumbrado	Problema de Alum	567906.562	6140068.57
139719	UNA PARTE DEL AF	2/1/2019	Finalizado	13/02/2019	Arbolado	Arbolado	Arboles o ramas c	570837.548	6142525.97
140217	Solicita que concurra	2/1/2017	Finalizado	3/1/2017	Limpieza	Contenedores	No paso Camion R	578094.825	6140949.75
158018	Hay un basural fuera	2/1/2018	Ingresado	2/1/2018	Limpieza	Problema de lim	Residuos fuera de	573377.853	6145610.98
158019	Ramas caidas sobre	2/1/2019	Finalizado	11/1/2019	CECOED	Emergencias	CECOED-Arbolado	575347.505	6142439.68
159617	Exp: 2016-3230-98-C	2/1/2017	En Proceso	2/1/2017	Saneamiento	Bocas de Torme	Boca de Tormenta	575505.23	6137254.09
159618	2 cuerpos Veterinaria	2/1/2018	Finalizado	4/1/2018	Limpieza	Problema de lim	Animal muerto der	573596.957	6143019.36
159619	Árbol inclinado a 45	2/1/2019	En Proceso	2/3/2019	Arbolado	Arbolado	Arbol deteriorado /	566218.096	6139533.61
160017	Contenedor repleto c	2/1/2017	En Proceso	2/1/2017	Limpieza	Contenedores	No paso Camion R	571108.518	6146611.09
107218	Se encuentran dos b	02/01/18	En Proceso	11/09/13	Calles y veredas	Viales	Bache	570307.493	6146981.7

Figura 8.7: Ejemplos de Datos de Reclamos para el análisis

Por otro lado, en la Figura 8.8 se muestran algunos procesos que se definieron en la herramienta DataCleaner para analizar estos datos. Notar que se utilizan algunos pasos de conversiones porque los datos se importan desde un formato CSV[1] y se cargan como campos de texto.

[1] comma separated values

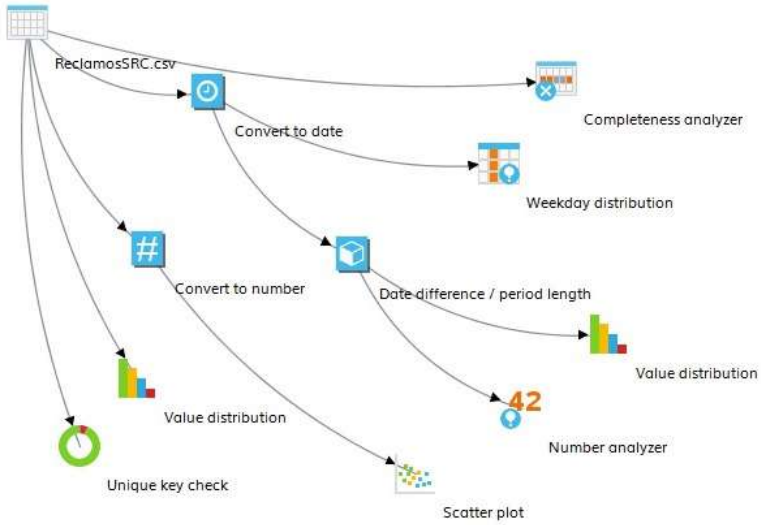


Figura 8.8: Procesos de Análisis de los Datos de Muestra

Ubicación de los Reclamos (DataCleaner)

Como se puede observar en la Figura 8.7, la ubicación de los reclamos se almacena en dos campos numéricos: CoordX y CoordY. La Figura 8.9 presenta el resultado de un análisis *Scatter Plot* utilizando estos campos, el cual permite ver la distribución de la ubicación de los reclamos.

Figura 8.9: Análisis de la Localización de los Reclamos por Ploteo

Aunque DataCleaner no tiene herramientas para trabajar con datos espaciales, utilizando esta forma de graficar se puede observar que los reclamos se distribuyen de una forma similar a la silueta del departamento de Montevideo. En particular, se detecta que un reclamo se encuentra alejado del resto y fuera de los límites de Montevideo.

Tabla 8.14: Resultados de *Scatter Plot* sobre Colección Base de Datos SRC IM

Resultados Técnica	
Técnica	*Scatter Plot*
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Tabla reclamos, atributos coordx-coordy	reclamos fuera de Montevideo

La aplicación de esta técnica permite detectar un nuevo problema de calidad de datos:

- P10: Existen reclamos cuya ubicación está fuera de los límites de Montevideo.

Unicidad de ID de Reclamos (DataCleaner)

Para verificar la unicidad de los identificadores de los reclamos, se utiliza una técnica de validación de unicidad clave (i.e. Unique Key Check). Como se muestra en la Figura 8.10, en este muestreo de datos no hay identificadores repetidos.

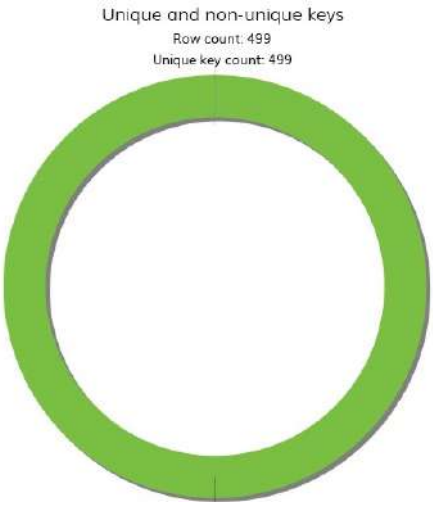
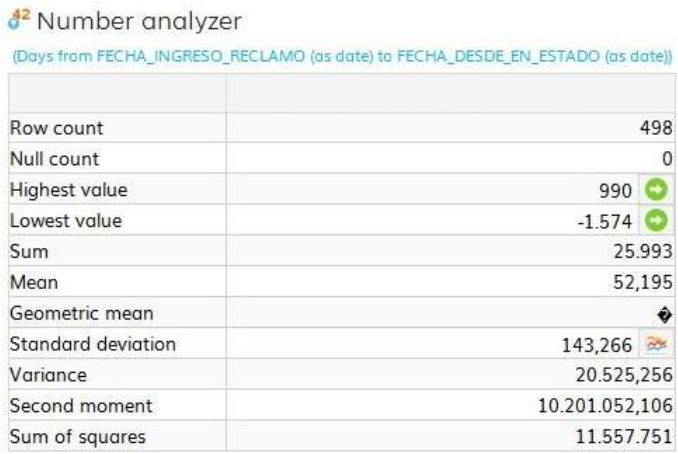


Figura 8.10: Análisis de la Unicidad de los Identificadores de Reclamos

Por lo tanto, con esta técnica no se detecta ningún nuevo problema de calidad de datos.

Correlación entre Fecha de Ingreso del Reclamo y Fecha de Último Estado (DataCleaner)

En este caso se plantea explorar con una técnica de correlación, la relación entre la fecha de ingreso del reclamo y la última fecha de cambio de estado. Para esto se aplica una conversión de datos, cálculo de la diferencia entre las fechas y un análisis numérico de dicha diferencia. Los resultados del análisis numérico se ven en la Figura 8.11.


Number analyzer

(Days from FECHA_INGRESO_RECLAMO (as date) to FECHA_DESDE_EN_ESTADO (as date))

Row count	498
Null count	0
Highest value	990
Lowest value	-1.574
Sum	25.993
Mean	52,195
Geometric mean	
Standard deviation	143,266
Variance	20.525,256
Second moment	10.201.052,106
Sum of squares	11.557.751

Figura 8.11: Diferencia entre Fecha Ingreso Reclamo y Fecha Último Cambio de Estado

Tabla 8.15: Resultados de Análisis Numérico sobre Colección Base de Datos SRC IM

Resultados Técnica	
Técnica	*Análisis Numérico*
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Atributos: Reclamos.fecha, EstadoReclamo.fechaEstado	Valor Mínimo = -1574, valor máximo: 990

La aplicación de esta técnica permite detectar dos nuevos problemas de calidad de datos:

- P11: Existen reclamos que tienen como fecha de último cambio de estado una fecha previa a su ingreso (diferencia entre fechas negativa)
- P12: El valor máximo encontrado para la diferencia entre las fechas (que brinda un indicador del tiempo que demora en resolverse un reclamo) es de 990 días.

Distribución de Valores de Áreas Reclamos

Por último, se utiliza la técnica de distribución de valores para analizar qué valores toman los atributos y cuáles son los más frecuentes. En general no se encuentran problemas de clasificación de los reclamos porque la mayoría de los datos son elegidos por los ciudadanos en base a opciones predefinidas (p. ej. categorías de reclamos).

Sin embargo, como se muestra en la Figura 8.12, se detecta que para el caso de un reclamo aparece el área Arbolado1 en lugar de Arbolado.

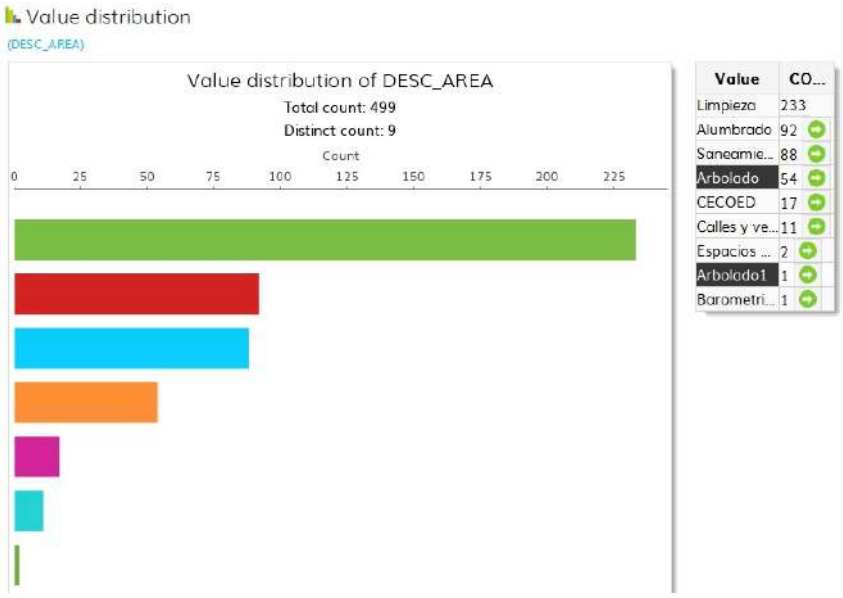


Figura 8.12: Distribución de Valores de Áreas de Atención de Reclamos

Tabla 8.16: Resultados de Distribución de Valores sobre Colección Base de Datos SRC IM

Resultados Técnica	
Técnica	*Distribución de Valores*
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Tablas: Reclamos, SubcategoriasReclamo	SubcategoriasReclamo.nombre = Arbolado1

La aplicación de esta técnica permite detectar un nuevo problema de calidad de datos:

- P13: Existen nombres de categorías que nos son correctos.

Resumen Nuevos Problemas de Calidad de Datos

La Tabla 8.17 presenta un resumen de los nuevos problemas de calidad de datos detectados utilizando técnicas de *Data Profiling*.

Tabla 8.17: Problemas Detectados con Técnicas de *Data Profiling*.

Resultados Técnicas	**Problemas**			
	P10	**P11**	**P12**	**P13**
Scatter Plot (1)	X			
Análisis Numérico (1)		X	X	
Distribución de Valores (1)				X

8.4.5. Definir Nuevos Requerimientos de Calidad de Datos

En esta actividad se definen nuevos requerimientos de calidad de datos en base a los problemas detectados en la actividad anterior. Los nuevos requerimientos identificados son:

- R13 - La localización del reclamo debe estar dentro de los límites de Montevideo.
- R14 - La fecha de cualquier cambio de estado del reclamo debe ser posterior a la fecha del reclamo.
- R15 - El cambio de estado de un reclamo a «resuelto» debería producirse dentro de las 48 horas posteriores a que el problema haya sido solucionado efectivamente.

La Tabla 8.18 presenta la relación entre los nuevos requerimientos que surgen a partir de los problemas detectados con técnicas de *Data Profiling*.

Tabla 8.18: Relaciones entre Problemas y Requerimientos de Calidad

Problemas	**Requerimientos**		
	R13	**R14**	**R15**
P10	X		
P11		X	
P12			X
P13			

9 Definir Estrategia de Gestión de Calidad

Este capítulo describe la tercera etapa del proceso de gestión de calidad, la cual consiste en definir la estrategia para llevar a cabo la gestión de la calidad de datos en el escenario. La definición de la estrategia se realiza en base a los resultados obtenidos en etapas anteriores y a prioridades que se establecen en esta etapa.

La Sección 9.1 presenta los objetivos y resultados esperados de la etapa y la Sección 9.2 describe el marco conceptual asociado a la misma. La Sección 9.3 detalla las actividades a realizar en la etapa y la Sección 9.4 describe estas actividades en el marco del caso de estudio presentado en el Capítulo 4.

9.1. Objetivos y Resultados Esperados

El objetivo de esta etapa es definir la estrategia para gestionar la calidad de datos en el escenario, la cual establece la forma de abordar los requerimientos de calidad identificados.

Los principales roles del CCD involucrados en la etapa Definir Estrategia de Gestión de Calidad son el Responsable de Calidad de Datos, el Experto de Negocio y el Analista de Calidad de Datos.

Los resultados esperados de esta etapa son:

- la priorización de los requerimientos de calidad de datos
- la definición de la estrategia de gestión de la calidad de datos en el escenario

9.2. Marco Conceptual Asociado

Esta sección describe los principales conceptos asociados a la etapa Definir Estrategia de Gestión de Calidad. Estos conceptos permiten especificar las prioridades de los requerimientos de calidad de datos identificados, así como describir la estrategia de gestión de calidad de datos para el escenario en base a pasos que abordan requerimientos específicos.

La Figura 9.1 presenta los conceptos asociados a estos elementos.

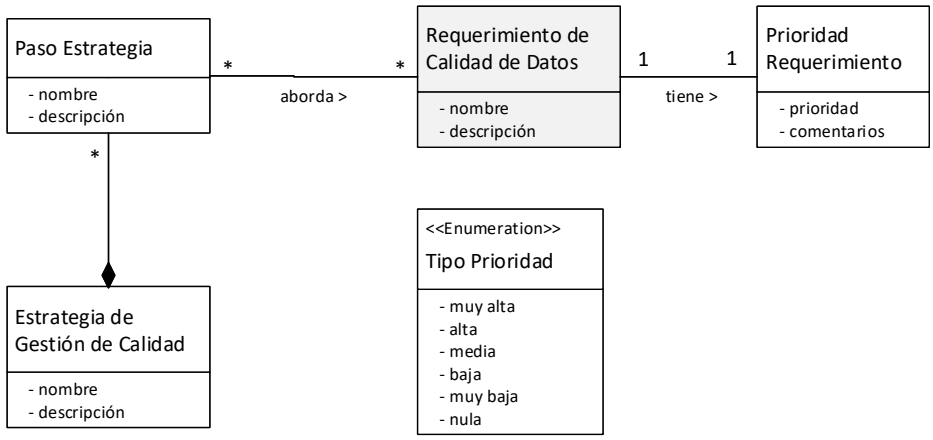


Figura 9.1: Estrategia de Gestión de Calidad y Prioridades de Requerimientos

9.3. Actividades de la Etapa

Esta sección detalla las actividades de la etapa Definir Estrategia, las cuales se presentan gráficamente en la Figura 9.2.

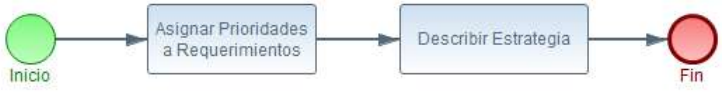


Figura 9.2: Actividades de Definir Estrategia

9.3.1. Asignar Prioridades a Requerimientos

La primera actividad de la etapa Definir Estrategia de Gestión de Calidad consiste en asignar prioridades a todos los requerimientos de calidad de datos identificados. Para esta asignación de prioridades se deben tener en cuenta los resultados de etapas anteriores que estén vinculados a los requerimientos:

- los problemas de los cuales surgen los requerimientos
- los actores interesados en los requerimientos
- los elementos del escenario (p. ej. colecciones) a los cuales aplican los requerimientos
- las estimaciones de cumplimiento de requerimientos
- las estimaciones del alcance de los problemas asociados a requerimientos

Se sugiere registrar la asignación de prioridades a requerimientos en una planilla con la estructura que se presenta en la Tabla 9.1.

Tabla 9.1: Asignación de Prioridades a Requerimientos

Requerimientos	**Prioridad**	**Comentarios**
Requerimiento 1	alta	Comentario 1
Requerimiento 2	media	Comentario 2
Requerimiento 3	media	Comentario 3

9.3.2. Describir Estrategia

La segunda actividad de la etapa Definir Estrategia de Gestión de Calidad consiste en describir la estrategia para la gestión de calidad en el escenario, en base a las prioridades de los requerimientos y a los recursos disponibles para abordarlos.

Se sugiere registrar la información de la estrategia en planillas con la estructura que se presenta en la Tabla 9.2 y la Tabla 9.3.

Tabla 9.2: Describir Estrategia

Estrategia de Gestión de Calidad de Datos	
Nombre Estrategia	Nombre 1
Descripción Estrategia	Descripción 1

Tabla 9.3: Pasos de la Estrategia

Requerimiento	**Paso 1**	**Paso 2**
Requerimiento 1	X	
Requerimiento 2		X
Requerimiento 3	X	

9.4. Aplicación en el Caso de Estudio

Esta sección presenta la etapa Definir Estrategia de Gestión de Calidad en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4), siguiendo las actividades definidas en la Sección 9.3.

9.4.1. Asignar Prioridades a Requerimientos

En este actividad se asignan prioridades a todos los requerimientos identificados en base a los resultados obtenidos en etapas anteriores. La Tabla 9.4 lista los requerimientos identificados en etapas anteriores y la Tabla 9.5 presenta la asignación de prioridades a estos requerimientos.

Tabla 9.4: Requerimientos Identificados

ID	Requerimiento
R1	El número de documento del ciudadano tiene que corresponder a una cédula válida.
R2	El email del ciudadano debe tener un formato de correo electrónico válido.
R3	Los nombres y apellidos de los ciudadanos deben ser los reales.
R4	El email del ciudadano debe existir.
R5	El domicilio del ciudadano debe existir.
R6	El teléfono del ciudadano debe existir
R7	La edad del ciudadano que se registra debe estar entre 10 y 110 años.
R8	El reclamo debe corresponder a un hecho real.
R9	El reclamo no debe estar duplicado.
R10	El estado «rechazado» de un reclamo debería tener una aclaración no vacía.
R11	La localización del reclamo debe ser consistente con su categoría.
R12	La categoría del reclamo debe ser consistente con su fecha.
R13	La localización del reclamo debe estar dentro de los límites de Montevideo.
R14	La fecha de cualquier cambio de estado del reclamo debe ser posterior a la fecha del reclamo.
R15	El cambio de estado de un reclamo a «resuelto» debería producirse dentro de las 48 horas posteriores a que el problema haya sido solucionado efectivamente.

A modo de ejemplo, al requerimiento R1 se le asignó prioridad «alta» dado que es de interés para un gran número de actores y aplica a varios elementos del escenario (p. ej. colecciones, clientes) que son también de interés para varios actores.

Tabla 9.5: Asignación de Prioridades a Requerimientos

Req.	Prioridad	Comentarios
R1	alta	
R2	media	
R3	baja	
R4	alta	
R5	alta	
R6	alta	
R7	baja	
R8	baja	
R9	alta	
R10	alta	
R11	media	
R12	baja	
R13	alta	
R14	alta	
R15	alta	

9.4.2. Describir Estrategia

En este actividad se define y describe la estrategia de gestión de calidad de datos, en base a la asignación de prioridades a requerimientos. En particular, para el caso de estudio se determina que la estrategia consista en abordar, en primer lugar, los requerimientos con prioridad alta y muy alta, para luego abordar los requerimientos con otras prioridades.

Tabla 9.6: Describir Estrategia

Estrategia de Gestión de Calidad de Datos	
Nombre Estrategia	Estrategia Gestión Calidad para Sistema SRC
Descripción Estrategia	Abordar en primer lugar los requerimientos con prioridad alta y muy alta. Luego abordar los requerimientos con otras prioridades.

De esta forma, la estrategia consiste de dos pasos:

1. Paso 1: se abordan los requerimientos con prioridad alta y muy alta
2. Paso 2: se aborda el resto de los requerimientos

La Tabla 9.7 presenta los requerimientos que se abordan en cada paso de la estrategia.

Tabla 9.7: Pasos de la Estrategia

Requerimiento	Paso 1	Paso 2
R1	X	
R2		X
R3		X
R4	X	
R5	X	
R6	X	
R7		X
R8		X
R9	X	
R10	X	
R11		X
R12		X
R13	X	
R14	X	
R15	X	

10 Definir Modelo de Calidad de Datos

La cuarta etapa del proceso consiste en la definición de un modelo de calidad de datos para el escenario de trabajo. Este modelo se construye en base al modelo de referencia presentado en el Capítulo 3 y a los requerimientos de calidad de datos incluidos en la estrategia definida en el Capítulo 9.

La Sección 10.1 presenta los objetivos y resultados esperados de esta etapa y la Sección 10.2 describe el marco conceptual asociado a la misma. La Sección 10.3 detalla las actividades a realizar en la etapa y la Sección 10.4 describe estas actividades en el marco del caso de estudio presentado en el Capítulo 4.

10.1. Objetivos y Resultados Esperados

El objetivo de esta etapa es definir el modelo de calidad de datos para el escenario de trabajo en el que se aplique el *framework*.

Los principales roles del CCD involucrados en la etapa Definir Modelo de Calidad de Datos son el Responsable de Calidad de Datos, el Experto de Negocio y el Analista de Calidad de Datos.

El resultado esperado de esta etapa es el modelo de calidad de datos para el escenario específico en que se aplique el *framework*, incluyendo:

- la identificación de los elementos del modelo de referencia a utilizar
- la definición de los elementos base del modelo
- la definición de perfiles de evaluación

10.2. Marco Conceptual Asociado

Esta sección describe los principales conceptos que, junto con los descriptos en la Sección 3.1, permiten especificar modelos de calidad de datos específicos para escenarios en los que se aplique el *framework*. En particular, la Figura 10.1 presenta los conceptos: Métrica Instanciada, Regla de Evaluación y Perfil de Evaluación.

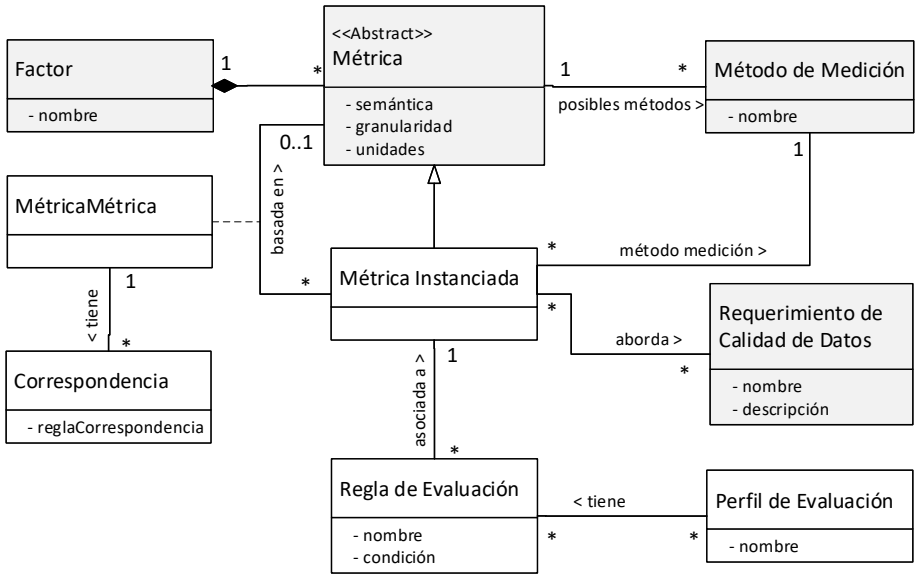


Figura 10.1: Métricas Instanciadas y Reglas de Evaluación

Una Métrica Instanciada es una métrica definida para ser utilizada en un escenario de aplicación concreto. Una métrica instanciada puede basarse en una métrica genérica o específica. En este caso es necesario especificar las correspondencias entre los elementos que maneja la métrica genérica o específica (p. ej. atributo país) con los elementos del escenario (p. ej. columna País de la tabla Persona de la base de datos Ciudadanos). Por último, una métrica instanciada aborda requerimientos de calidad de datos y tiene asociado un método de medición, que será el que se utilice para tomar medidas de esa métrica.

Por otro lado, una Regla de Evaluación especifica una condición que permite evaluar si el resultado de una medición para una métrica instanciada se encuentra (o no) dentro de los rangos de valores esperados. Las reglas de evaluación se pueden agrupar en Perfiles de Evaluación. El Ejemplo 49 presenta la definición de tres perfiles de evaluación, cada uno de los cuales tiene una única regla asociada a una métrica, pero con distintos rangos de valores esperados.

Ejemplo 49

Nombre: Perfil Básico
Regla de Evaluación: Regla Básica
Métrica Instanciada: RatioNoNulos-DireccionCliente
Condición: resultado > 50 %

Nombre: Perfil Intermedio
Regla de Evaluación: Regla Intermedia
Métrica Instanciada: RatioNoNulos-DireccionCliente
Condición: resultado > 70 %

Nombre: Perfil Avanzado
Regla de Evaluación: Regla Avanzada
Métrica Instanciada: RatioNoNulos-DireccionCliente
Condición: resultado > 90 %

10.3. Actividades de la Etapa

Esta sección detalla las actividades de la etapa Definir Modelo de Calidad de Datos, las cuales se presentan gráficamente en la Figura 10.2.

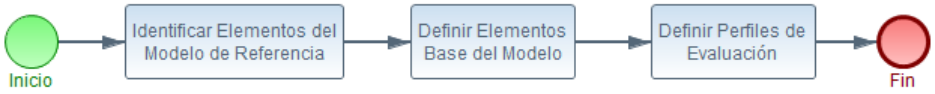


Figura 10.2: Actividades de Definir Modelo de Calidad

10.3.1. Identificar Elementos del Modelo de Referencia

La primera actividad de la etapa Definir Modelo de Calidad de Datos consiste en identificar los elementos del modelo de referencia a utilizar para construir el modelo. Para esto se deben considerar todos los requerimientos de calidad incluidos en la estrategia definida.

Para identificar un primer conjunto de factores a incluir en el modelo, se sugiere utilizar la Tabla 10.1, la cual presenta dimensiones y factores del modelo de referencia asociados a problemas / requerimientos de calidad de datos comunes.

Tabla 10.1: Problemas / Requerimientos de Calidad y Factores de Calidad

Problema	Requerimiento	Factor
números de teléfono que le faltan dígitos	los números de teléfono deben tener 8 dígitos	Correctitud Sintáctica
errores de estandarización	los códigos para los países deben seguir el estándar ISO3166	Correctitud Sintáctica
una dirección está dada por un departamento (p. ej. Colonia) y no por la calle y número	todas las direcciones deben tener calle y número de puerta	Precisión
números no tienen decimales	todos los números deben tener 2 dígitos después de la coma	Precisión
solo una parte de los clientes están registrados en la base de datos de clientes	la base de datos debe tener al menos el 80 % de los clientes	Cobertura
datos que no son vigentes	los datos deben haber sido actualizados como máximo hace un mes	Actualidad
la fecha de nacimiento es de un niño pero la edad es 40	la fecha de nacimiento y la edad deben ser consistentes	Integridad intra-relación
violación de claves foráneas	la base de datos debe satisfacer todas las claves foráneas	Integridad inter-relación
valores fuera de rango	el rango debe ser (1,10)	Integridad de Dominio
entidades aparecen repetidas	el 90 % de los datos no deben estar duplicados	Duplicación
una entidad aparece repetida con contradicciones	el 90 % de los datos no deben tener contradicciones	Contradicción

Luego de identificados los factores, se puede analizar si en el modelo de referencia existen métricas (en dichos factores) adecuadas para los requerimientos del escenario.

Se sugiere registrar la información de los elementos del modelo de referencia a utilizar en una planilla con la estructura que se presenta en la Tabla 10.2, indicando además de qué requerimiento del escenario surgen.

Tabla 10.2: Requerimientos / Elementos del Modelo de Referencia

Requerimiento	Dimensión	Factor	Métrica
RQ1	Dimensión 1	Factor 1	Métrica 1
RQ2	Dimensión 2	Factor 2	Métrica 2
RQ3	Dimensión 3	Factor 3	Métrica 3

Cabe recalcar que el modelo de calidad de datos para el escenario puede incluir también dimensiones, factores o métricas que no estén definidos en el modelo de referencia. En este caso se sugiere analizar la conveniencia de extender el modelo de referencia con estos nuevos elementos.

10.3.2. Definir Elementos Base del Modelo

La segunda actividad de la etapa Definir Modelo de Calidad de Datos consiste en definir los elementos base del modelo (i.e. dimensiones, factores y métricas) tomando como base los resultados de la actividad anterior.

En particular, se debe definir al menos una métrica para cada uno de los requerimientos a abordar. Como se presentó en la Sección 3.1.2, una nueva métrica puede:

- estar basada en una métrica genérica
- estar basada en una métrica específica
- definirse desde cero

Se sugiere registrar la información de la definición de cada métrica en una planilla con la estructura que se presenta en la Tabla 10.3.

Tabla 10.3: Definición de una Métrica

Definición Métrica	
Dimensión:	Dimensión 1
Factor:	Factor 1
Basada en:	Métrica
Nombre:	Nombre
Semántica:	Semántica
Granularidad:	Granularidad
Tipo Resultado:	Tipo Resultado
Reglas Correspondencia:	Reglas Correspondencia

10.3.3. Definir Perfiles de Evaluación

La tercera actividad de la etapa Definir Modelo de Calidad de Datos consiste en definir perfiles de evaluación en base a reglas de evaluación para las métricas definidas y considerando los requerimientos de calidad identificados.

Se sugiere registrar la información de los perfiles de evaluación en una planilla con la estructura que se presenta en la Tabla 10.4.

Tabla 10.4: Perfil de Evaluación

Perfil de Evaluación: Perfil 1		
Regla	**Métrica**	**Condición**
Regla de Evaluación 1	métrica 1	condición 1
Regla de Evaluación 2	métrica 2	condición 2
Regla de Evaluación 3	métrica 3	condición 3
Regla de Evaluación 4	métrica 4	condición 4

10.4. Aplicación en el Caso de Estudio

Esta sección presenta la etapa Definir Modelo de Calidad de Datos en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 10.3.

10.4.1. Identificar Elementos del Modelo de Referencia

En esta actividad se identifican los elementos del modelo de referencia a utilizar. Para esto, se toma como base los tres primeros requerimientos incluidos en el primer paso de la estrategia (cf. Sección 9.4.2) y se identifican elementos (i.e. dimensiones, factores y métricas) del modelo de referencia que estén asociados.

La Tabla 10.5 presenta los tres primeros requerimientos incluidos en el primer paso de la estrategia

Tabla 10.5: Requerimientos Identificados

ID	Requerimiento
R1	El número de documento del ciudadano tiene que corresponder a una cédula válida.
R4	El email del ciudadano debe existir.
R5	El domicilio del ciudadano debe existir.

La Tabla 10.6 presenta los elementos del modelo de referencia asociados a estos requerimientos.

Tabla 10.6: Requerimientos de Calidad y Elementos del Modelo de Referencia

Req.	Dimensión	Factor	Métrica Genérica / Específica
R1	Exactitud	Correctitud Sintáctica	Formato (NumeroDocumento, DNIC)
R4	Exactitud	Correctitud Semántica	ErrorCorrectitudSemDebil
R5	Exactitud	Correctitud Semántica	ErrorCorrectitudSemDebil

10.4.2. Definir Elementos Base del Modelo

En esta actividad se definen los elementos del modelo de calidad de datos a construir, tomando como insumo los elementos del modelo de referencia identificados en la actividad anterior.

De esta forma, el modelo de calidad de datos para el escenario incluirá, en principio, la dimensión Exactitud y los factores identificados en la Tabla 10.6.

Con respecto a las métricas, a continuación se analiza cómo definir una métrica instanciada para cada uno de los requerimientos a abordar.

El requerimiento R1 implica medir la correctitud sintáctica en documentos de identidad uruguayos. El modelo de referencia posee una métrica específica para dicha situación, por lo que se genera en el modelo una métrica instanciada M1 basada en dicha métrica específica. La métrica M1 se define en la Tabla 10.7.

Tabla 10.7: Definición de la Métrica M1

Métrica M1	
Dimensión:	Exactitud
Factor:	Correctitud Sintáctica
Basada en:	Formato (NumeroDocumento, DNIC)
Nombre:	Formato (numDoc, DNIC)
Semántica:	Indica si el valor del atributo numDoc cumple con el formato de cédula de identidad uruguaya establecido por DNIC, que establece que la cédula tiene siete dígitos seguidos de un octavo dígito verificador que está en función de los otros siete $(d_1d_2d_3d_4d_5d_6d_7 - v, v = f(d_i))$
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	NumeroDocumento=Ciudadano.numDoc Estandar(NumeroDocumento)=«DNIC»

El requerimiento R4 implica medir la correctitud semántica en direcciones de correo electrónico. El modelo de referencia no posee una métrica específica para dicha situación, pero sí posee una métrica genérica, por lo que se genera en el modelo una métrica instanciada M4 basada en dicha métrica genérica. La métrica M4 se define en la Tabla 10.8.

Tabla 10.8: Definición de la Métrica M4

Métrica M4	
Dimensión:	Exactitud
Factor:	Correctitud Semántica
Basada en:	CorrectitudSemDebil
Nombre:	CorrectitudSemDebil(eMail)
Semántica:	Evalúa si una dirección de correo electrónico existe. Se utiliza una función en un lenguaje de alto nivel que permite verificar la existencia a través del envío de un correo electrónico y el análisis de la respuesta.
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	Atr=Ciudadano.eMail Función= existeEmail(eMail: String): Boolean

El requerimiento R5 implica medir la correctitud semántica en domicilios ingresados como texto libre. El modelo de referencia no posee una métrica específica para dicha situación, pero sí posee una métrica genérica, por lo que se genera en el modelo una métrica instanciada M5 basada en dicha métrica genérica. La métrica M5 se define en la Tabla 10.9.

Tabla 10.9: Definición de la Métrica M5

Métrica M5	
Dimensión:	Exactitud
Factor:	Correctitud Semántica
Basada en:	CorrectitudSemDebil
Nombre:	CorrectitudSemDebil(domicilio)
Semántica:	Evalúa si una dirección geográfica existe. Se utiliza una capa geográfica de direcciones de Montevideo como referencial.
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	Atr=Ciudadano.domicilio Referencial= DireccionesMontevideo

Por último, para cada una de las métricas definidas se establece un posible método para su medición. La Tabla 10.10 presenta una descripción de estos métodos.

Tabla 10.10: Métodos de Calidad de Datos del Modelo

Métrica	**Método**	**Descripción Método**
M1	Met1	Se ejecuta una rutina en el lenguaje Java que verifica el largo correcto del número de documento y compara el dígito verificador almacenado con el valor calculado.
M4	Met4	Se ejecuta una rutina en el lenguaje Java que envía un correo electrónico a cada dirección, utilizando el protocolo SMTP, y luego analiza la respuesta para comprobar si la dirección no existe (códigos 551 y 554).
M5	Met5	Se ejecuta una rutina en el lenguaje Java que normaliza y geocodifica la dirección ingresa, utilizando capas de direcciones de Montevideo dadas por nombre de calle y número de puerta. Las direcciones que no pueden ser normalizadas debido a que provienen de un atributo de texto libre, no necesariamente puede considerarse que no existen, por lo que el resultado es una aproximación con cierto margen de error.

10.4.3. Definir Perfiles de Evaluación

En esta actividad se definen perfiles de evaluación en base a reglas de evaluación para las métricas instanciadas que resultaron de la actividad anterior. En particular, para cada una de las métricas presentadas en la Sección 10.4.2 con resultado de tipo Boolean, se define una regla de evaluación que tiene como condición que el resultado sea «true».

La Tabla 10.11 presenta la definición de un perfil de evaluación con varias de estas reglas de evaluación.

Tabla 10.11: Perfil de Evaluación

Perfil de Evaluación: Perfil Básico		
Regla	**Métrica**	**Condición**
RegDoc: Cumple Formato Documento	M1	resultado = true
RegEmail: Existe Email	M4	resultado = true
RegDomic: Existe Dirección Geo	M5	resultado = true

11 Medir y Evaluar la Calidad de Datos

La quinta etapa del proceso consiste en la medición y evaluación de la calidad de datos en el escenario de trabajo, tomando como base el modelo de calidad definido en la etapa anterior.

La Sección 11.1 presenta los objetivos y resultados esperados de esta etapa y la Sección 11.2 describe el marco conceptual asociado a la misma. La Sección 11.3 detalla las actividades a realizar en la etapa y la Sección 11.4 describe estas actividades en el marco del caso de estudio presentado en el Capítulo 4.

11.1. Objetivos y Resultados Esperados

El objetivo de esta etapa es realizar mediciones utilizando las métricas y métodos definidos en el modelo de calidad, así como evaluar la calidad de los datos en base a los perfiles y reglas de evaluación, también definidos en el modelo.

Los principales roles del CCD involucrados en la etapa Medir y Evaluar Calidad de Datos son el Responsable de Calidad de Datos, el Técnico de Calidad de Datos y el Analista de Calidad de Datos.

Los resultados esperados de esta etapa son:

- implementación de métodos de medición para las métricas que lo requieran
- medidas para las distintas métricas definidas en el modelo
- evaluación de la calidad de datos en base a estas medidas así como a los perfiles y reglas de evaluación

11.2. Marco Conceptual Asociado

Esta sección describe los principales conceptos asociados a la medición y evaluación de la calidad de datos. En particular, la Figura 11.1 presenta los conceptos: Medida, Evaluación Medida, Evaluación Regla, Evaluación Perfil y Objeto Medible.

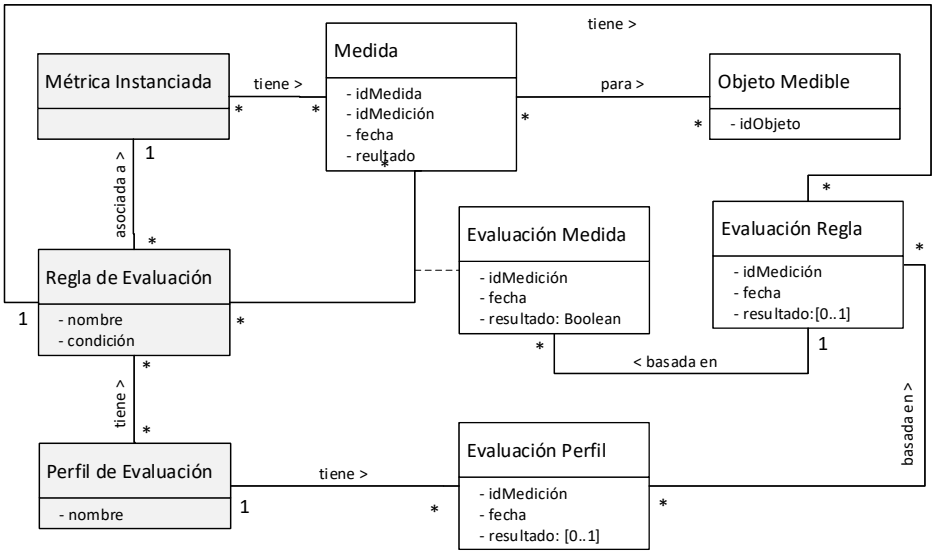


Figura 11.1: Medición y Evaluación de la Calidad de Datos

Una Medida representa el resultado de realizar una medición para una métrica instanciada, utilizando el método de medición asociado. Las medidas aplican a uno o más objetos medibles, cuyo tipo depende de la granularidad de la métrica instanciada. Por ejemplo, si la granularidad de la métrica es «instanciaEntidad», el objeto medible para las medidas de esa métrica es una instancia de una entidad (p. ej. un ciudadano dado). Cabe recalcar que idMedida identifica una medida, mientras idMedición agrupa un conjunto de medidas realizadas en un mismo proceso de medición.

La forma de identificar un objeto medible (i.e. idObjeto) depende del tipo de objeto. Por ejemplo, si el objeto medible refiere a una organización, basta con especificar el nombre de la organización. Por otro lado, si el objeto medible es una instancia de una entidad, se debe especificar el nombre de la organización, la colección de datos donde se almacena la entidad, el nombre de la entidad y el identificador de la entidad (p. ej. el valor de la clave primaria si los datos de la entidad se almacenan en una tabla de una base de datos relacional).

Una Evaluación Medida representa el resultado de la evaluación de una medida en base a una regla de evaluación. El resultado de la evaluación de una medida es de tipo «Boolean» e indica si la condición de la regla se cumple o no en base al resultado de la medida. Por otro lado, una Evaluación Regla representa el resultado de la evaluación de una regla de evaluación en base a las evaluaciones de las medidas con respecto a esa regla. El resultado de la evaluación de un regla es un valor entre 0 y 1, que indica la proporción de evaluaciones de medidas con esa regla cuyo resultado es «true».

Por otro lado, una Evaluación Perfil representa el resultado de la evaluación de un perfil de evaluación en base a las evaluaciones de las reglas asociadas al perfil. El resultado de la evaluación de un perfil es un valor entre 0 y 1 calculado de acuerdo a la siguiente fórmula (donde r es el resultado de la evaluación de cada regla del perfil y n es la cantidad de reglas del perfil):

$$\sum_{i=1}^{n} r_i/n$$

11.3. Actividades de la Etapa

Esta sección detalla las actividades de la etapa Medir y Evaluar Calidad de Datos, las cuales se presentan gráficamente en la Figura 11.2.

Figura 11.2: Actividades de Medir y Evaluar Calidad de Datos

11.3.1. Implementar Métodos de Medición

La primera actividad de la etapa Medir y Evaluar Calidad de Datos consiste en implementar métodos de medición para las métricas que así lo requieran.

Esta actividad es necesaria dado que si bien en la etapa de definición del modelo (c.f. Capítulo 10) se especifican los métodos a utilizar para cada métrica, puede ocurrir que los métodos no estén implementados o requieran adaptaciones.

11.3.2. Medir Calidad de Datos

La segunda actividad de la etapa Medir y Evaluar Calidad de Datos consiste ejecutar los métodos de medición, de forma tal de tomar medidas para las métricas identificadas.

Se sugiere almacenar las medidas resultantes de cada medición en una estructura como la que se presenta en la Tabla 11.1.

Tabla 11.1: Medidas para Métrica

idMedida	idMedición	Fecha	Resultado	idObjeto
1	1234	2019-12-02	resultado 1	id objeto 1
2	1234	2019-12-02	resultado 2	id objeto 2
3	1234	2019-12-02	resultado 3	id objeto 3
4	1235	2019-12-04	resultado 4	id objeto 1
5	1235	2019-12-04	resultado 5	id objeto 2
6	1235	2019-12-04	resultado 6	id objeto 3

11.3.3. Evaluar Calidad de Datos

La tercera actividad de la etapa Medir y Evaluar Calidad de Datos consiste evaluar la calidad de datos en base a las medidas tomadas así como a las reglas y perfiles de evaluación. Para esto se debe, en primera instancia, evaluar las medidas tomadas para cada métrica en base a cada regla. En segunda instancia, se debe evaluar cada regla en base a las evaluaciones de las medidas utilizando esa regla. Por último, se debe evaluar cada perfil de evaluación en base a las evaluaciones de las reglas asociadas al perfil.

Se sugiere almacenar la evaluación de las medidas en una estructura como la que se presenta en la Tabla 11.2.

Tabla 11.2: Evaluaciones Medidas

Regla	Métrica	idMedida	idMedición	Fecha	Resultado
R1	M1	1	1234	2019-12-02	false
R1	M1	2	1234	2019-12-02	true
R1	M1	3	1234	2019-12-02	true
R1	M1	4	1235	2019-12-04	true
R1	M1	5	1235	2019-12-04	true
R1	M1	6	1235	2019-12-04	true

Por otro lado, se sugiere almacenar la evaluación de las reglas en una estructura como la que se presenta en la Tabla 11.3.

Tabla 11.3: Evaluación Reglas

idMedición	Regla	Fecha	Resultado
1234	Regla 1	2019-12-02	0,33
1234	Regla 2	2019-12-02	0,66
1235	Regla 1	2019-12-04	0,5
1235	Regla 2	2019-12-04	1

Por otro lado, se sugiere almacenar la evaluación de los perfiles en una estructura como la que se presenta en la Tabla 11.4.

Tabla 11.4: Evaluación Perfiles

idMedición	Perfil	Fecha	Resultado
1234	Perfil 1	2019-12-02	0,66
1234	Perfil 2	2019-12-02	0,5
1235	Perfil 1	2019-12-04	1
1235	Perfil 2	2019-12-04	1

11.4. Aplicación en el Caso de Estudio

Esta sección presenta la etapa Medir y Evaluar Calidad de Datos en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 11.3. En particular, se considera que hay definido un único perfil con las reglas definidas en la Tabla 10.11.

11.4.1. Implementar Métodos de Medición

En esta actividad se implementan o adaptan métodos de medición para poder tomar medidas de calidad de datos en el escenario de trabajo dado.

En este caso, se asume que ya se contaba con una implementación base de los métodos requeridos de acuerdo a lo especificado en la Tabla 10.10. Por este motivo, para el caso de estudio esta etapa sólo implica la adaptación de los métodos de medición de forma que puedan acceder a las colecciones de datos para tomar las medidas (p. ej. especificando una cadena de conexión para que puedan acceder una tabla de una base de datos relacional).

11.4.2. Medir Calidad de Datos

En esta actividad, se ejecutan los métodos de medición para tomar las medidas de calidad de datos, de acuerdo a las métricas identificadas, y se almacenan dichas medidas para su posterior evaluación.

La Tabla 11.5 presenta ejemplos de medidas para la métrica «M1: Formato (numDoc, DNIC)» identificada en la Sección 10.4.2.

La Tabla 11.6 presenta ejemplos de medidas para la métrica «M4: CorrectitudSemDebil (email)» identificada en la Sección 10.4.2.

La Tabla 11.7 presenta ejemplos de medidas para la métrica «M5: CorrectitudSemDebil (domicilio)» identificada en la Sección 10.4.2.

Tabla 11.5: Medidas Métrica M1: Formato (numDoc, DNIC)

idMedida	idMedición	Fecha	Resultado	idOrganización	idColección	idEntidad	idAtributo	idObj
1	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	numDoc	25
2	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	numDoc	27
3	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	numDoc	28
4	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	numDoc	30

Tabla 11.6: Medidas MétricaM4: CorrectitudSemDebil (email)

idMedida	idMedición	Fecha	Resultado	idOrganización	idColección	idEntidad	idAtributo	idObj
10	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	email	25
11	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	email	27
12	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	email	28
13	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	email	30

Tabla 11.7: Medidas Métrica M5: CorrectitudSemDebil (domicilio)

idMedida	idMedición	Fecha	Resultado	idOrganización	idColección	idEntidad	idAtributo	idObj
21	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	domicilio	25
22	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	domicilio	27
23	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	domicilio	28
24	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	domicilio	30

11.4.3. Evaluar Calidad de Datos

En esta actividad, se evalúa la calidad de datos en el escenario de trabajo en base a las medidas tomadas así como a las reglas y perfiles de evaluación.

La Tabla 11.8 presenta la evaluación de las distintas medidas tomadas en base a las reglas definidas para el caso de estudio, cuya condición establecía que el resultado de la medida debía tener un valor «true».

Tabla 11.8: Evaluaciones Medidas

Regla	Métrica	idMedida	idMedición	Fecha	Resultado
RegDoc	M1	1	1234	2019-12-02	true
RegDoc	M1	2	1234	2019-12-02	true
RegDoc	M1	3	1234	2019-12-02	false
RegDoc	M1	4	1234	2019-12-02	true
RegEmail	M4	10	1234	2019-12-02	false
RegEmail	M4	11	1234	2019-12-02	true
RegEmail	M4	12	1234	2019-12-02	true
RegEmail	M4	13	1234	2019-12-02	false
RegDomic	M5	21	1234	2019-12-02	false
RegDomic	M5	22	1234	2019-12-02	false
RegDomic	M5	23	1234	2019-12-02	true
RegDomic	M5	24	1234	2019-12-02	true

La Tabla 11.9 presenta la evaluación de las reglas definidas para el caso de estudio, en base a la evaluación de las medidas realizada con dichas reglas.

Tabla 11.9: Evaluación Reglas - Caso de Estudio

idMedición	Regla	Fecha	Resultado
1234	RegDoc	2019-12-02	0,75
1234	RegEmail	2019-12-02	0,5
1234	RegDomic	2019-12-04	0,5

La Tabla 11.10 presenta la evaluación del perfil definido para el caso de estudio, en base a la evaluación de las reglas asociadas al perfil.

Tabla 11.10: Evaluación Perfiles

idMedición	Perfil	Fecha	Resultado
1234	Perfil Básico	2019-12-02	0,58

12 Determinar Causas Problemas

La sexta etapa del proceso consiste en determinar las causas de los problemas de calidad de datos que se confirmen o detecten en base a la evaluación de la calidad.

La Sección 12.1 presenta los objetivos y resultados esperados de esta etapa y la Sección 12.2 describe el marco conceptual asociado a la misma. La Sección 12.3 detalla las actividades a realizar en la etapa y la Sección 12.4 describe estas actividades en el marco del caso de estudio presentado en el Capítulo 4.

12.1. Objetivos y Resultados Esperados

El objetivo de esta etapa es determinar las causas de los problemas de calidad de datos que se confirmen o detecten en base a la evaluación de la calidad de datos.

Los principales roles del CCD involucrados en la etapa Determinar Causas de Problemas de Calidad son el Responsable de Calidad de Datos, el Técnico de Calidad de Datos y el Analista de Calidad de Datos.

Los resultados esperados de esta etapa son:

- confirmación de problemas de calidad resultantes de etapas anteriores, en base a la evaluación de la calidad de datos
- nuevos problemas de calidad, detectados en base a la evaluación de la calidad de datos
- causas de los problemas de calidad de datos, determinadas en base a resultados de etapas anteriores (p. ej. caracterización técnica y de negocio del escenario)

12.2. Marco Conceptual Asociado

Esta sección describe los principales conceptos asociados a las causas de problemas de calidad de datos. En particular, la Figura 12.1 presenta los conceptos: Causa de Problema de Calidad y Tipo de Causa de Problema.

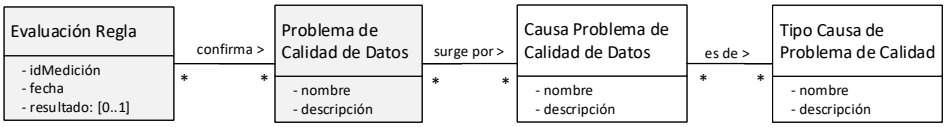


Figura 12.1: Problemas de Calidad de Datos y Causas

Los problemas de calidad de datos pueden confirmarse o detectarse a partir de evaluaciones de reglas.

Una Causa de Problema de Calidad representa un motivo por el cual se pueden originar problemas de calidad. Por ejemplo, el hecho de que no se realicen validaciones del formato de fechas en un sistema, puede ser uno de los motivos por el cual las fechas ingresadas a través de ese sistema no se estén almacenando en un mismo formato.

Un Tipo de Causa de Problema representa una familia de causas de problemas. La Tabla 12.1 presenta un conjunto inicial de tipos de problemas que maneja el *framework*, el cual puede ser extendido de acuerdo a nuevos requerimientos.

Tabla 12.1: Tipos de Causas de Problemas

Nombre
Defecto en aplicación
Defecto en proceso de datos
Capacitación usuarios
Problemas en procedimientos
Falta de restricciones en colecciones de datos

12.3. Actividades de la Etapa

Esta sección detalla las actividades de la etapa Determinar Causas de Problemas de Calidad, las cuales se presentan gráficamente en la Figura 12.2.

Figura 12.2: Actividades de Determinar Causas de Problemas de Calidad

12.3.1. Confirmar y Detectar Problemas de Calidad de Datos

La primera actividad de la etapa Determinar Causas de Problemas de Calidad consiste en confirmar y detectar problemas de calidad, en base a los resultados de la evaluación de la calidad de datos realizada en la etapa anterior.

Para confirmar los problemas de calidad de datos resultantes de etapas anteriores, se sugiere analizar si estos problemas son confirmados (o no) por las evaluaciones de las reglas asociadas a las métricas que abordan los requerimientos originados por estos problemas.

Para detectar nuevos problemas de calidad de datos se sugiere, en primer lugar, analizar el nivel de cumplimiento de los perfiles, comenzando por los de menores niveles y determinando cuáles son las reglas de evaluación que los originan. En segundo lugar, se sugiere analizar si existen problemas de calidad que surjan a partir de estas reglas con bajos niveles de cumplimiento y que no hayan sido identificados en etapas anteriores.

Se sugiere registrar esta información en una planilla con la estructura que se presenta en la Tabla 12.2.

Tabla 12.2: Confirmación y Detección de Problemas de Calidad de Datos

Problema	Nombre	Reglas	idMedición
P1	Problema 1	R1, R2	id1
P2	Problema 2	R2	id2
P3	Problema 3	R3	id3

12.3.2. Determinar Causas de Problemas

La segunda actividad de la etapa consiste en determinar las causas de los problemas de calidad de datos, en base a los resultados de las etapas anteriores.

Para esto se sugiere analizar, para cada problema, la forma en que los datos a los que hace referencia el problema fluyen a través de los clientes de datos y se almacenan en las colecciones. Los datos a los que hace referencia el problema quedan determinados por los objetos medibles de las medidas evaluadas por reglas cuya evaluación confirma o detecta el problema.

Como resultado de este análisis se apunta a obtener un conjunto de causas candidatas para cada problema, las cuales posiblemente requieran un análisis adicional (p. ej. que involucre a otros roles del CCD) para confirmarlas.

Se sugiere almacenar las causas confirmadas de los problemas en una planilla con la estructura que se presenta en la Tabla 12.3.

Tabla 12.3: Causas Problemas de Calidad

Problema	Causa	Descripción Causa	Tipo Causa
P1	C1	Descripción Causa 1	Tipo Causa 1
P2	C2	Descripción Causa 2	Tipo Causa 1
P3	C3	Descripción Causa 3	Tipo Causa 2

12.4. Aplicación en el Caso de Estudio

Esta sección presenta la etapa Determinar Causas de Problemas de Calidad en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 12.3.

12.4.1. Confirmar y Detectar Problemas de Calidad de Datos

En esta actividad se confirman y detectan problemas de calidad en base a los resultados de la evaluación de la calidad de datos.

Para confirmar los problemas ya identificados para el caso de estudio, se analizan las evaluaciones de las reglas vinculadas, en particular, para los siguientes problemas:

- P1: Correos electrónicos inexistentes
- P2: Domicilios no encontrados

La regla vinculada al problema P1 es RegEmail. Por lo que se puede observar en la Tabla 11.9, la evaluación de esta regla resulta en un valor de 0,55 por lo que se entiende que el problema queda confirmado por esta evaluación.

La regla vinculada al problema P2 es RegDomic. De forma similar, la evaluación de esta regla (cf. Tabla 11.9) tiene un valor de 0,5, por lo que también se entiende que el problema P2 queda confirmado por la misma.

Por otro lado, se puede observar que la evaluación de la regla RegDoc incluida en el Perfil Básico es de 0,75. Esto permite detectar un nuevo problema para el caso de estudio:

- P13: Documentos con formato no válido

La Tabla 12.4 resume este análisis presentando las evaluaciones de qué reglas confirman o detectan los problemas de calidad de datos considerados previamente.

Tabla 12.4: Confirmación y Detección de Problemas de Calidad para Caso de Estudio

Problema	Nombre	Reglas	idMedición
P1	Correos electrónicos inexistentes	RegEmail	1234
P2	Domicilios no encontrados	RegDomic	1234
P13	Documentos con formato no válido	RegDoc	1234

12.4.2. Determinar Causas de Problemas de Calidad

En esta actividad se determinan las causas de los problemas de calidad confirmados en el marco del caso de estudio.

Para esto se analizan las formas en que los datos asociados a los tres problemas considerados (documento de identidad, email y domicilio) fluyen a través de los clientes de datos y son almacenados en las colecciones.

En particular, en la caracterización técnica y de negocio del escenario se puede observar que estos datos son ingresados por los ciudadanos mediante la aplicación móvil y fluyen a través de los servicios web y la aplicación empresarial, que es la que los almacena en la base de datos. Las causas de los problemas de calidad pueden estar asociadas entonces a cualquiera de estos componentes así como a procedimientos generales de la organización.

La Tabla 12.5 presenta las causas que se determinan y confirman en el marco del caso de estudio.

Tabla 12.5: Causas Problemas de Calidad - Caso de Estudio

Problema	Causa	Descripción Causa	Tipo Causa
P1	Falta mecanismo verificación email	Cuando el ciudadano especifica un email no existe un mecanismo que controle que el email existe (p. ej. enviando un correo de confirmación).	Defecto en aplicación
P1	Falta mecanismo actualización email	Si bien el email pudo haber sido especificado en forma correcta por el ciudadano, no existe un procedimiento que asegure que el email se mantenga actualizado (p. ej. para el caso de que el ciudadano no utilice más un email)	Problemas en procedimientos
P2	Falta validación de dirección	Cuando el ciudadano especifica una dirección, no se valida que la misma exista	Defecto en aplicación
P13	Falta validación documento	Cuando el ciudadano especifica un documento, no se valida que su formato sea correcto	Defecto en aplicación

13 Definir, Ejecutar y Evaluar Plan Mejora

La séptima etapa del proceso consiste en definir, ejecutar y evaluar un plan de mejora para la calidad de datos en el escenario de trabajo.

La Sección 13.1 presenta los objetivos y resultados esperados de esta etapa y la Sección 13.2 describe el marco conceptual asociado a la misma. La Sección 13.3 detalla las actividades a realizar en la etapa y la Sección 13.4 describe estas actividades en el marco del caso de estudio presentado en el Capítulo 4.

13.1. Objetivos y Resultados Esperados

El objetivo de esta etapa es definir, ejecutar y evaluar un plan de mejora para la calidad de datos en el escenario de trabajo.

Los principales roles del CCD involucrados en la etapa Determinar Causas de Problemas de Calidad son el Responsable de Calidad de Datos, el Técnico de Calidad de Datos y el Analista de Calidad de Datos.

Los resultados esperados de esta etapa son:

- definición de un plan de mejora de la calidad de datos
- resultados de la ejecución y evaluación del plan de mejora

13.2. Marco Conceptual Asociado

Esta sección describe los principales conceptos asociados al plan de mejora. En particular, la Figura 13.1 presenta los conceptos: Plan de Mejora, Acción de Mejora y Estrategia de Mejora.

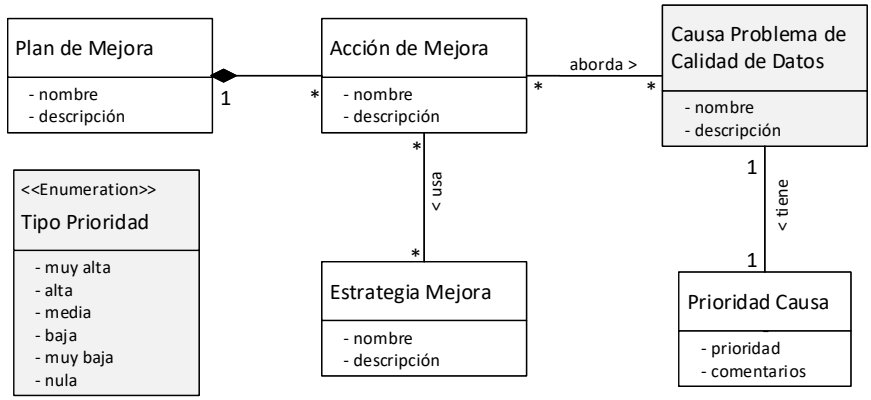


Figura 13.1: Plan de Mejora Calidad de Datos

Un Plan de Mejora es un conjunto de acciones que tienen como fin abordar las causas de los problemas de calidad detectados en un escenario de trabajo. Cada Acción de Mejora puede abordar varias causas de problemas de calidad, los cuales tienen una cierta prioridad, y utilizar distintas estrategias.

Una Estrategia de Mejora (p. ej. limpieza de datos, re-estructura de procedimientos) especifica una forma de abordar las causas y problemas de calidad de datos existentes en los escenarios. La Tabla 13.1 presenta un conjunto inicial de estrategias de mejora, las cuales pueden ser extendida.

Tabla 13.1: Estrategias de Mejora

Nombre
Limpieza de Datos
Re-estructura de Procedimientos
Corrección de Defectos

13.3. Actividades de la Etapa

Esta sección detalla las actividades de la etapa Definir, Ejecutar y Evaluar Plan de Mejora, las cuales se presentan gráficamente en la Figura 13.2.

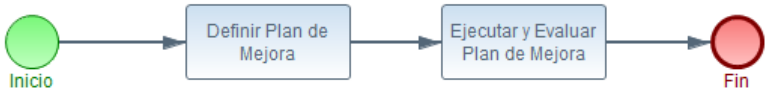


Figura 13.2: Actividades de Definir, Ejecutar y Evaluar Plan de Mejora

13.3.1. Definir Plan de Mejora

La primera actividad de esta etapa consiste en definir un plan de mejora para la calidad de datos en el escenario de trabajo.

Dado que las tareas de mejora son en general costosas (no solo desde el punto de vista económico), en esta actividad es importante identificar los problemas de mayor importancia para la organización así como los costos asociados para resolverlos. Para esto es recomendable crear una lista de causas de problemas a abordar, ordenada por prioridad en base a la importancia y costo, la cual se puede construir teniendo en cuenta distintos criterios:

- Dar más importancia a la estrategia de negocio de la organización:
  - organización que está apuntando al marketing directo, debería priorizar las causas de los problemas en los datos de clientes
  - organización que está enfocada a mejorar eficiencia de operaciones, debería priorizar las causas de los problemas en datos logísticos
- Asociación con problemas del negocio ya conocidos (p. ej. reuniones perdidas con clientes: direcciones incorrectas)
- Tasas de errores reales vs. requerimientos de nivel de calidad
- Económicas, ya que hay causas que generan problemas que tienen consecuencias más costosas que otras

Se sugiere registrar esta información en una planilla con la estructura que se presenta en la Tabla 13.2.

Tabla 13.2: Prioridades Causas Problemas de Calidad

Problema	Causa	Prioridad
Problema 1	Causa 1	prioridad 1
Problema 1	Causa 2	prioridad 2
Problema 2	Causa 3	prioridad 1

Una vez que se definan las causas a abordar con el plan de mejora, es necesario definir las acciones a tomar así como las estrategias a utilizar para cada caso. A modo de ejemplo, la estrategia de limpieza de datos suele ser más adecuada para conjuntos de datos estáticos, mientras que las estrategias de re-estructuración de procedimientos y corrección de defectos suelen ser mejores para los conjuntos de datos que están en constante cambio (por ejemplo, creciendo constantemente).

Por otro lado, la estrategia a utilizar también depende del tipo de causa que se aborde (p. ej. la estrategia corregir defecto de aplicación es aplicable para causas que refieren a defectos de una aplicación).

Se sugiere registrar las acciones de un plan de mejora en una planilla con la estructura que se presenta en la Tabla 13.3.

Tabla 13.3: Acciones de Plan de Mejora

Acción	Descripción	Estrategias	Causas
Acción 1	Descripción 1	Estrategia 1, Estrategia 2	Causa 1, Causa 2
Acción 2	Descripción 2	Estrategia 1	Causa 3

13.3.2. Ejecutar y Evaluar Plan de Mejora

La segunda actividad de esta etapa consiste en ejecutar y evaluar el plan de mejora.

Para evaluar el plan de mejora se sugiere realizar una nueva medición de la calidad de datos en el escenario y determinar en cuánto mejoró el nivel de cumplimiento de los perfiles de evaluación considerados.

13.4. Aplicación en el Caso de Estudio

Esta sección presenta la etapa Definir, Ejecutar y Evaluar Plan de Mejora en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 13.3.

13.4.1. Definir Plan de Mejora

En esta actividad se define el plan de mejora para el escenario de trabajo en base a un conjunto de acciones de mejora y tomando como base las prioridades de las causas de los problemas de calidad determinados.

La Tabla 13.4 presenta la lista de prioridades para las distintas causas de problemas de calidad de datos determinados para el caso de estudio.

Por otro lado, la Tabla 13.5 presenta las acciones a tomar en el marco del plan de mejora.

Tabla 13.4: Prioridades - Causas Problemas de Calidad - Caso de Estudio

Problema	Causa	Prioridad
P1	Falta mecanismo verificación email	alta
P1	Falta mecanismo actualización email	alta
P2	Falta validación de dirección	muy alta
P13	Falta validación documento	alta

Tabla 13.5: Acciones de Plan de Mejora - Caso de Estudio

Acción	Descripción	Estrategias	Causas
A1	Se debe actualizar la aplicación para que incluya una funcionalidad de verificación de correo electrónico	Corrección defecto	Falta mecanismo verificación email
A2	Se debe establecer un procedimiento (manual o automático) que permita mantener actualizadas las direcciones de correo electrónico	Re-estructura procedimiento	Falta mecanismo actualización email
A3	Se debe actualizar la aplicación para que verifique una dirección al momento que se ingrese	Corrección defecto	Falta validación de dirección
A4	Se debe actualizar la aplicación para que verifique el formato del documento de identidad	Corrección defecto	Falta validación documento

13.4.2. Ejecutar y Evaluar Plan de Mejora

En esta actividad se ejecuta y evalúa el plan de mejora en base a la evaluaciones de los perfiles. A modo de ejemplo, se considera que luego de ejecutar el plan de mejora se obtuvo el resultado que se presenta en la Tabla 13.6 para la evaluación del perfil básico.

Tabla 13.6: Evaluación Perfiles - Plan de Mejora

idMedición	Perfil	Fecha	Resultado
1234	Perfil Básico	2019-12-29	0,80

De esta forma, el nivel de cumplimiento del perfil aumentó en más de un 30 %.

14 Recursos de Soporte

Este capítulo incluye recursos de soporte para la aplicación del *framework*.

14.1. Servicios de la Plataforma de Interoperabilidad

Servicio Básico de Información. Servicio provisto por la Dirección Nacional de Identificación Civil (DNIC) que devuelve los datos filiatorios asociados a un número de cédula.
https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/servicio-basico-informacion

Servicio de Direcciones. Servicio provisto por la Infraestructura de Datos Espaciales de Uruguay (IDEuy), el Correo Uruguayo y AGESIC que permite realizar búsquedas direcciones, localidades y sugerencias de calles de todo el territorio nacional.
https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/servicio-direcciones

Salud.uy - Servicios Terminológicos. Servicio provisto por Salud.uy que permite acceder al servidor de terminología.
https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/saluduy-servicios-terminologicos

Consulta de Entidad por RUT. Servicio provisto por la Dirección General de Impositiva (DGI) que dado un número de RUT lo valida y retorna la razón social y domicilio fiscal asociados.
https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/consulta-entidad-rut

14.2. Herramientas de Soporte

Talend Open Studio.
https://www.talend.com/products/talend-open-studio/

Data Cleaner.
https://datacleaner.org/

14.3. Estándares

[ISO15] ISO/IEC. *ISO/IEC 25024:2015 - Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality*. Estándar. International Organization for Standardization (ISO), oct. de 2015. URL: https://www.iso.org/standard/35749.html

[ISO13] ISO. *ISO 19157:2013 - Geographic information – Data quality*. Estándar. International Organization for Standardization (ISO), dic. de 2013. URL: https://www.iso.org/standard/32575.html

[ISO11] ISO. *ISO/TS 8000-1:2011*. Estándar. International Organization for Standardization (ISO), dic. de 2011. URL: https://www.iso.org/standard/50798.html

[ISO08] ISO/IEC. *ISO/IEC 25012:2008 - Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model*. Estándar. International Organization for Standardization (ISO), dic. de 2008. URL: https://www.iso.org/standard/35736.html

14.4. Fuentes de Consulta

14.4.1. Libros de Referencia

[BS16] Carlo Batini y Monica Scannapieco. *Data and Information Quality*. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-24106-7

[Pia18] Mario Piattini, Ismael Caballero, Ana Gómez y Fernando Gualo. *Calidad de Datos*. RA-MA EDITORIAL, 2018. ISBN: 978-84-9964-750-0

[Int17] DAMA International. *DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition*. Technics Publications, 2017. ISBN: 978-1634622349

14.4.2. Artículos Académicos

[Cab08] Ismael Caballero, Angélica Caro, Coral Calero y Mario Piattini. «IQM3: Information Quality Management Maturity Model». En: *Journal of Universal Computer Science* 14.22 (dic. de 2008), págs. 3658-3685

[Tep17] Jaak Tepandi, Mihkel Lauk, Janar Linros, Priit Raspel, Gunnar Piho, Ingrid Pappel y Dirk Draheim. «The Data Quality Framework for the Estonian Public Sector and Its Evaluation». En: *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXV*. Ed. por Abdelkader Hameurlain, Josef Küng, Roland Wagner, Sherif Sakr, Imran Razzak y Alshammari Riyad. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017, págs. 1-26. ISBN: 978-3-662-56121-8. DOI: 10.1007/978-3-662-56121-8_1

[Pul16] Venkata Sai Venkatesh Pulla, Cihan Varol y Murat Al. «Open Source Data Quality Tools: Revisited». En: *Information Technology: New Generations*. Ed. por Shahram Latifi. Cham: Springer International Publishing, 2016, págs. 893-902. ISBN: 978-3-319-32467-8

14.4.3. Reportes Técnicos de la Industria

[MA19] Melody Chien y Ankush Jain. *Magic Quadrant for Data Quality Tools*. Inf. téc. Gartner, mar. de 2019

[Hea18] Health Information and Quality Authority. *Background paper to support guidance for a data quality framework*. en. Inf. téc. Oct. de 2018, pág. 86

15 Aplicación del Framework

En este capítulo se describe la aplicación del *framework* al caso de estudio del Capítulo 4, siguiendo el Proceso de Gestión de Calidad de Datos definido en el Capítulo 5. En particular, el capítulo presenta de forma unificada la aplicación del *framework* descripta entre el Capítulo 6 y el Capítulo 13, incorporando además aspectos adicionales del caso de estudio tanto del soporte a la gestión de reclamos como a la toma de decisiones.

15.1. Caracterización Técnica y de Negocio

Esta sección presenta la caracterización técnica y de negocio del escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4), siguiendo las actividades definidas en la Sección 6.3. En particular, se extiende lo presentado en la sección 6.4 agregando elementos del soporte a la toma de decisiones (p. ej. varias intendencias, Presidencia y ministerios).

15.1.1. Representar Gráficamente el Escenario

En esta actividad se representa gráficamente el escenario de trabajo planteado. En este caso se aprovecha la descripción del escenario presentada en la Figura 4.3 y en la Figura 4.4 de la Sección 4.2. La Figura 15.1 y la Figura 15.2 muestran esta representación gráfica.

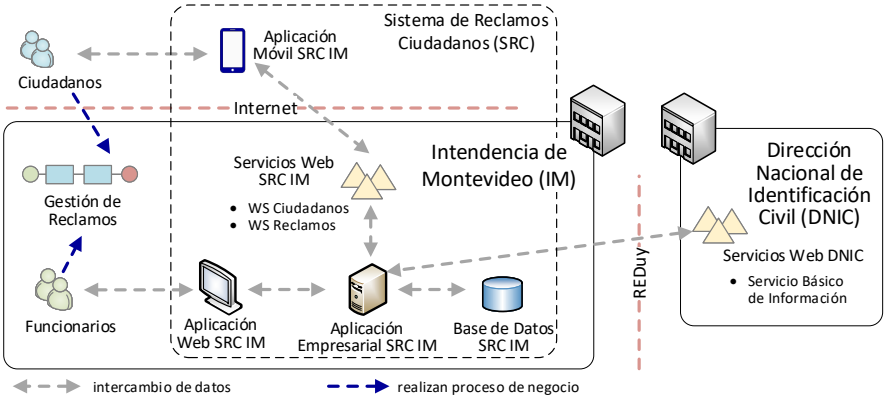


Figura 15.1: Representación Gráfica del Escenario del Caso de Estudio - Parte 1

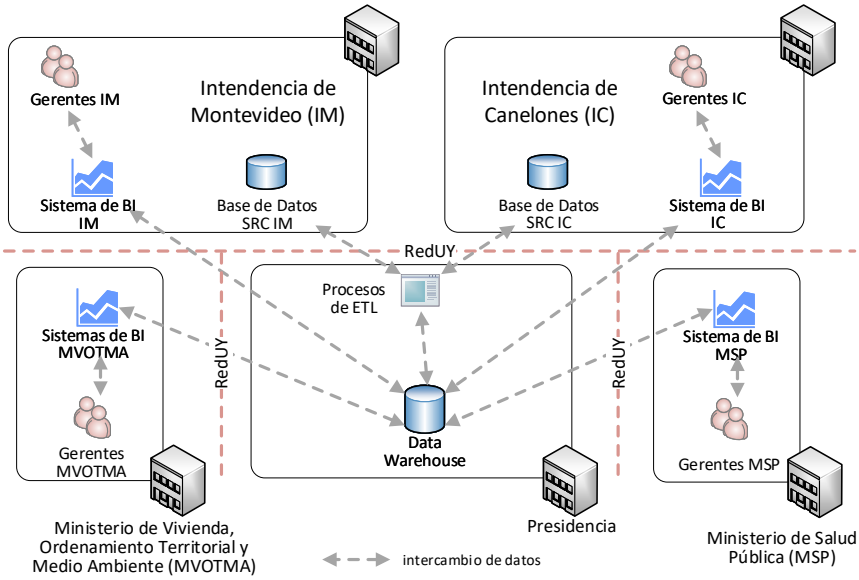


Figura 15.2: Representación Gráfica del Escenario del Caso de Estudio - Parte 2

15.1.2. Identificar Elementos de Negocio y Técnicos del Escenario

En esta actividad se identifican los elementos de negocio y técnicos del escenario, utilizando la representación gráfica elaborada en la actividad anterior. La Tabla 15.1 presenta estos elementos.

Tabla 15.1: Elementos de Negocio y Técnicos del Escenario de Trabajo.

Conceptos	**Elementos Identificados**
Organizaciones	Intendencia de Montevideo (IM), Intendencia de Canelones (IC), Dirección Nacional de Identificación Civil (DNIC), Ministerio de Salud Pública (MSP), Presidencia, Ministerio de Vivienda, Ordenamiento Territorial y Medio Ambiente (MVOTMA)
Dominio de Aplicación	Ciudadanía, Servicios Públicos
Colecciones de Datos	Base de Datos SRC IM, Base de Datos SRC IC, Data Warehouse
Clientes de Datos	
Servicios	WS Ciudadanos IM, WS Reclamos IM, WS Ciudadanos IC, WS Reclamos IC, Servicio Básico de Información
Procesos	Proceso de ETL
Aplicaciones	Aplicación Móvil SRC IM, Aplicación Web SRC IM, Aplicación Empresarial SRC IM, Aplicación Móvil SRC IC, Aplicación Web SRC IC, Aplicación Empresarial SRC IC
Sistemas	Sistema de BI IM, Sistema de BI IC
Roles de Usuarios	Ciudadano, Funcionario IM, Gerente IM, Funcionario IC, Gerente IC
Entidades de Negocio	Ciudadano, Reclamo
Procesos de Negocio	Gestión de Reclamos
Actores de Datos	Ciudadanos Zonas Costeras, Funcionarios IM Ventanilla, Funcionarios IM Áreas, Gerentes IM, Sección Evaluación y Monitoreo IM, Funcionarios IC Ventanilla, Funcionarios IC Áreas, Gerentes IC, Sección Evaluación y Monitoreo IC, Gerentes de Ministerios, Gerentes de Presidencia

15.1.3. Identificar Relaciones entre Clientes de Datos

En esta actividad se identifican las relaciones entre clientes de datos. La Tabla 15.2 presenta las relaciones identificadas. En particular, se presentan los clientes de datos asociados a una de las intendencias y los que brindan soporte a la toma de decisiones.

Tabla 15.2: Relaciones entre Clientes de Datos.

	WS Ciudadanos IM	**WS Reclamos IM**	**Servicio Básico de Información**	**Proceso de ETL**	**Aplicación Móvil SRC IM**	**Aplicación Web SRC IM**	**Aplicación Empresarial SRC IM**	**Sistema de BI IM**	**Ciudadano**	**Funcionario IM**	**Gerente IM**
WS Ciudadanos IM							X				
WS Reclamos IM							X				
Servicio Básico de Información											
Proceso de ETL											
Aplicación Móvil SRC IM	X	X									
Aplicación Web SRC IM							X				
Aplicación Empresarial SRC IM			X								
Sistema de BI IM											
Ciudadano					X						
Funcionario IM						X					
Gerente IM								X			

15.1.4. Relaciones entre Organizaciones y Clientes / Colecciones de Datos

En esta actividad se identifican las relaciones entre organizaciones y clientes / colecciones de datos. La Tabla 15.3 presenta las relaciones entre organizaciones y clientes de datos, mientras que la Tabla 15.4 presenta las relaciones entre organizaciones y colecciones de datos 15.4.

Tabla 15.3: Relaciones entre Clientes de Datos y Organizaciones

Clientes de Datos	**Organizaciones**		
	IM	**DNIC**	**Presidencia**
WS Ciudadanos IM	responsable		
WS Reclamos IM	responsable		
Servicio Básico de Información	utiliza	responsable	
Proceso de ETL			responsable
Aplicación Móvil SRC IM	responsable		
Aplicación Web SRC IM	responsable, utiliza		
Aplicación Empresarial SRC IM	responsable, utiliza		
Sistema de BI IM	responsable, utiliza		
Ciudadano			
Funcionario IM	responsable		
Gerente IM	responsable		

Tabla 15.4: Relaciones entre Organizaciones y Colecciones de Datos

Colecciones de Datos	**Organizaciones**		
	IM	**DNIC**	**Presidencia**
Base de Datos SRC IM	responsable		
Data Warehouse	utiliza		responsable

15.1.5. Identificar Relaciones entre Clientes y Colecciones de Datos

En esta actividad se identifican las relaciones entre clientes y colecciones de datos. La Tabla 15.5 presenta estas relaciones (A: automática, M: manual).

Tabla 15.5: Relaciones entre Clientes y Colecciones de Datos

Clientes de Datos	**Colecciones**			
	Base de Datos SRC IM		**Data Warehouse**	
	Ciudadano	**Reclamo**	**Ciudadano**	**Reclamo**
WS Ciudadanos IM				
WS Reclamos IM				
Servicio Básico de Información				
Proceso ETL	consulta (A)	consulta (A)	alta (A) modifica (A)	alta (A) modifica (A)
Aplicación Móvil SRC IM				
Aplicación Web SRC IM				
Aplicación Empresarial SRC IM	alta (M) baja (M) modifica (M) consulta (M)	alta (M) baja (M) modifica (M) consulta (M)		
Sistema de BI IM			consulta (M)	consulta (M)
Ciudadano				
Funcionario IM				
Gerente IM				

15.1.6. Identificar Relaciones de Actores y Procesos

En esta actividad se identifican las relaciones de actores de datos y procesos de negocio. La Tabla 15.6 presenta las relaciones de actores de datos (U:utiliza, G: gestiona, I:interesa). En la Tabla 15.7 se presentan las relaciones de los procesos de negocio.

Tabla 15.6: Relaciones de Actores de Datos

Colecciones y Clientes de Datos	**Actores de Datos**						
	Ciudadanos Zonas Costeras	**Funcionarios IM Ventanilla**	**Funcionarios IM Áreas**	**Gerentes IM**	**Sección Evaluación y Monitoreo IM**	**Gerentes de Ministerios**	**Gerentes de Presidencia**
Colecciones de Datos							
Base de Datos SRC IM	G	G	G	I	U	I	I
Data Warehouse	I	I	I	U	I	I	I
Clientes de Datos							
WS Ciudadanos IM							
WS Reclamos IM							
Servicio Básico de Información							
Proceso de ETL							
Aplicación Móvil SRC IM	U			I			
Aplicación Web SRC IM		U	U	I			
Aplicación Empresarial SRC IM							
Sistema de BI IM				U	I	I	G
Ciudadano				I	I	I	I
Funcionario IM				I			I
Gerente IM				I			I

Tabla 15.7: Relaciones de Procesos de Negocio

	Procesos de Negocio
	Gestión de Reclamos
Clientes de Datos	Aplicación Web SRC IM Aplicación Empresarial SRC IM WS Ciudadanos IM WS Reclamos IM Servicio Básico de Información Proceso ETL Aplicación Móvil SRC IM Sistema BI IM Funcionario Funcionario IM Gerente IM
Colecciones de Datos	Base de Datos SRC IM
Organizaciones	IM, DNIC
Actores de Datos	Ciudadanos Zonas Costeras (utiliza, interesa) Funcionarios IM Ventanilla (utiliza, interesa) Funcionarios IM Áreas (utiliza, interesa) Gerentes IM (participa, interesa) Sección Evaluación y Monitoreo IM (interesa)
Entidades de Negocio	Reclamo, Ciudadano

15.2. Caracterización de Calidad de Datos

Esta sección presenta la caracterización de calidad del escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 7.3.

15.2.1. Identificar Requerimientos de Calidad de Datos

En esta actividad se identifican requerimientos de calidad de datos relevantes para el escenario de trabajo. La Tabla 15.8 presenta los requerimientos identificados.

Tabla 15.8: Requerimientos de Calidad de Datos

ID	**Requerimiento**
R1	El número de documento del ciudadano tiene que corresponder a una cédula válida.
R2	El email del ciudadano debe tener un formato de correo electrónico válido.

Por otro lado, la Tabla 15.9 presenta el interés de los actores de datos en los requerimientos identificados.

Tabla 15.9: Interés de Actores en Requerimientos de Calidad de Datos

Actores	**Requerimientos**	
	R1	**R2**
Ciudadanos Zonas Costeras		
Funcionarios IM Ventanilla	X	X
Funcionarios IM Áreas		X
Gerentes IM	X	X
Sección Evaluación y Monitoreo IM	X	X

15.2.2. Relaciones entre Requerimientos y Elementos del Escenario

En esta actividad se identifican los elementos del escenario a los cuales aplican los requerimientos resultantes de la actividad anterior. La Tabla 15.10 presenta las relaciones entre requerimientos y elementos del escenario.

Tabla 15.10: Relaciones entre Requerimientos de Calidad y Elementos del Escenario

Elementos	**Requerimientos**	
	R1	**R2**
Entidades de Negocio		
Ciudadano	X	X
Reclamo		
Atributos		
Ciudadano.nombres		
Ciudadano.apellidos		
Ciudadano.numDoc	X	
Ciudadano.email		X
Reclamo.id		
Reclamo.motivo		
Reclamo.fecha		
Reclamo.coordX		
Reclamo.coordY		
Colecciones de Datos		
Base de Datos SRC IM	X	X
Clientes de Datos		
WS Ciudadanos IM	X	X
WS Reclamos IM		
Servicio Básico de Información	X	
Proceso Gestión de Reclamos	X	X
Aplicación Móvil SRC IM	X	X
Aplicación Web SRC IM	X	X
Aplicación Empresarial SRC IM	X	X
Ciudadano	X	X
Funcionario IM	X	X
Organizaciones		
IM	X	X
DNIC	X	X
Procesos de Negocio		
Gestión de Reclamos	X	X

15.2.3. Identificar Problemas de Calidad de Datos

En esta actividad se identifican los problemas de calidad de datos relevantes para el escenario. En base al listado inicial de aspectos de calidad de datos presentado en la Sección 4.5, se elabora la planilla que se presenta en la Tabla 15.11.

Tabla 15.11: Problemas de Calidad de Datos

ID	Problemas
P1	**Nombres irreales.** Se han detectado nombres de fantasía, apodos o *nicknames* en los campos destinados a los nombres y apellidos del ciudadano.
P2	**Correos electrónicos inexistentes.** A algunos ciudadanos registrados no les llegan los correos electrónicos que envía el sistema porque ingresaron incorrectamente su dirección de correo.
P3	**Domicilios no encontrados.** Algunos de los domicilios ingresados por los ciudadanos no se corresponden con ninguna dirección oficial de la capa de direcciones que maneja el sistema.
P5	**Edades poco confiables.** Algunas edades de los ciudadanos registrados (que se calculan en base a la fecha de nacimiento ingresada) resultan poco confiables, por estar fuera de los rangos en que deberían encontrarse los usuarios de esta aplicación (mayores de 10 años y menores de 100 años).
P5	**Uso abusivo.** Se han constatado varios casos de usuarios que realizan un uso abusivo del sistema, que incluyen: reclamos de incidentes falsos, múltiples reclamos del mismo usuario sobre el mismo incidente, observaciones o fotos de los reclamos con contenido inapropiado o irrelevante.
P6	**Reclamos duplicados.** Una parte importante del trabajo del funcionario que recibe los reclamos es poder identificar los reclamos de distintos ciudadanos que hacen referencia al mismo problema. Si estos reclamos duplicados no son detectados oportunamente, puede suceder que las áreas reciban reclamos que ya fueron resueltos en base a otros reclamos.
P7	**Reclamos rechazados sin aclaraciones.** Algunos funcionarios no completan las observaciones cuando cambian el estado del reclamo. Esto genera disconformidad en algunos ciudadanos, que ven sus reclamos en estado «rechazado» y no conocen el motivo.
P8	**Inconsistencia entre la categoría del reclamo y su ubicación geográfica.** Se ha constado que a veces la categoría del reclamo no es compatible con su ubicación (p. ej. reclamo de la subcategoría «presencia de cianobacterias en playa» es reportado en una ubicación alejada de la playa).
P9	**Inconsistencia entre la categoría del reclamo y su fecha.** En algunos casos, la fecha en la que se ingresa el reclamo no es compatible con la categoría del reclamo (p. ej. se ingresa un reclamo de la subcategoría «falta de limpieza posterior a feria» en una fecha muy posterior a la realización de dicha feria).

Por otro lado, la planilla de la Tabla 15.12 especifica qué actores identificaron cada problema.

Tabla 15.12: Problemas de Calidad identificados por Actores

Actores	**Problemas identificados**								
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**
Actores de Datos									
Ciudadanos Zonas Costeras							X		
Funcionarios IM Ventanilla	X	X		X	X			X	X
Funcionarios IM Áreas			X			X			
Gerentes IM				X		X	X		
Sección Evaluación y Monitoreo							X	X	X

15.2.4. Definir Nuevos Requerimientos de Calidad de Datos

En esta actividad se definen nuevos requerimientos de calidad de datos en base a los problemas identificados en la actividad anterior. La Tabla 15.13 presenta la definición de estos nuevos requerimientos.

Tabla 15.13: Requerimientos de Calidad de Datos

ID	**Requerimiento**
R1	El número de documento del ciudadano tiene que corresponder a una cédula válida.
R2	El email del ciudadano debe tener un formato de correo electrónico válido.
R3	Los nombres y apellidos de los ciudadanos deben ser los reales.
R4	El email del ciudadano debe existir.
R5	El domicilio del ciudadano debe existir.
R6	El teléfono del ciudadano debe existir.
R7	La edad del ciudadano que se registra debe estar entre 10 y 110 años.
R8	El reclamo debe corresponder a un hecho real.
R9	El reclamo no debe estar duplicado.
R10	El estado «rechazado» de un reclamo debería tener una aclaración no vacía.
R11	La localización del reclamo debe ser consistente con su categoría.
R12	La categoría del reclamo debe ser consistente con su fecha.

Por otro lado, en la Tabla 15.14 se especifica el interés de los actores en estos nuevos requerimientos.

Por último, en la Tabla 15.15 se especifica de qué problemas surgieron los nuevos requerimientos.

Tabla 15.14: Interés de Actores en Requerimientos de Calidad de Datos

Actores	Requerimientos											
	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**
Ciudadanos Zonas Costeras										X		
Funcionarios IM Ventanilla	X	X	X	X			X	X			X	X
Funcionarios IM Áreas		X			X				X			
Gerentes IM	X	X					X		X	X		
Sección Evaluación y Monitoreo IM	X	X								X	X	X

Tabla 15.15: Relaciones entre Problemas y Requerimientos de Calidad

Problemas	Requerimientos											
	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R11**
P1			X									
P2				X								
P3					X							
P4							X					
P5								X				
P6									X			
P7										X		
P8											X	
P9												X

15.3. Examinar Datos Objetivo

Esta sección presenta la etapa Examinar Datos Objetivos en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4), siguiendo las actividades definidas en la Sección 8.3.

15.3.1. Clasificar Datos Objetivo

En esta actividad se clasifican los datos relevantes para el escenario, especificando los tipo de las colecciones y el tipo de datos de los atributos de las entidades de negocio.

En la Tabla 15.16 se detalla el tipo de la colección de datos «Base de Datos SRC IM» y en la Tabla 15.17 se muestran los tipos de datos de los atributos más relevantes para los requerimientos definidos para el escenario.

Tabla 15.16: Tipos de las Colecciones de Datos

Colección de Datos	Tipo de Colección
Base de Datos SRC IM	BD Relacional
Data Warehouse	BD Relacional

Tabla 15.17: Tipos de Dato de los Atributos de Ciudadano y Reclamo

Atributo	Tipo de Datos
Ciudadano.nombres	Alfanumérico
Ciudadano.apellidos	Alfanumérico
Ciudadano.numDoc	Alfanumérico
Ciudadano.email	Alfanumérico
Reclamo.id	Numérico
Reclamo.motivo	Alfanumérico
Reclamo.fecha	Fecha
Reclamo.coordX	Numérico
Reclamo.coordY	Numérico

15.3.2. Estimar Alcance de Problemas

En esta actividad se estima el alcance de los problemas de calidad identificados, aplicando técnicas de *Data Profiling*. En particular, se realiza una exploración directa de los datos de la colección «Base de Datos SRC IM» para estimar el alcance de los problemas que refieren a los nombres (i.e. P1), email (i.e. P2) y fecha de nacimiento de los ciudadanos (i.e. P3). La Figura 15.3 presenta algunos de estos datos.

ID	nombres	apellidos	numDoc	eMail	telefono	sexo	fechaNac
2	Juan	Perez	3.457.2158	jperez@gmal.com	91234543	M	03/07/1993
3	Felipe	Rodriguez	4.254.7516	ferrodri@homail.com	95457812	M	12/05/1905
1	García		5.384.2634	fvgarcia@montevideo.com.uy	93265498	F	10/08/2001
5	Ana González		1.234.5678	anita@adinet.com.uy	94569218	F	26/09/1975
10		Alejandra	5.945.6546	alej0234@adine.com.uy	93234658		02/08/2010
8		Jorge	4.569.0731	jorgito_25@gmail.com	99875093	M	26/02/2000
7	elcarbone		31684564	carboen@hotmal.com	98769054		25/04/2019
12		vidatrico	45439802	trico987@adinet.com	93456239	F	15/5/1903

Figura 15.3: Datos de Ciudadanos

Esta exploración permite estimar el alcance de los problemas antes mencionados:

- P1: Hay valores nulos y valores extraños en los campos de nombre y apellido de los ciudadanos (p. ej. elcarbone, vidatrico)
- P2: Hay valores en el atributo mail que aunque son casillas bien formadas, tienen errores en los dominios (p. ej. @hotmal.com, @gmal.com)

- P4: Hay valores extraños en las fechas de nacimiento (i.e. años 1903, 1906 y 2019)

La Tabla 15.18 y la Tabla 15.19 presentan cómo se registran estos resultados.

Tabla 15.18: Resultados de Exploración Directa sobre Colección Base de Datos SRC IM

Resultados Técnica	
Técnica	Exploración Directa
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Tabla ciudadanos, atributo nombres, apellidos	elcarbone, vidatrico
2. Tabla ciudadanos, atributo email	@hotmal.com, @gmal.com
3. Tabla ciudadanos, atributo fechaNac	1903, 1906, 2019

Tabla 15.19: Aplicación *Data Profiling* para Problemas de Calidad de Datos

Resultados Técnicas	**Problemas**		
	P1 (nombres)	**P2 (email)**	**P4 (fechaNac)**
Exploración Directa (1)	alto		
Exploración Directa (2)		alto	
Exploración Directa (3)			alto

15.3.3. Estimar Cumplimiento de Requerimientos

En esta actividad se estima el cumplimiento de los requerimientos de calidad identificados, aplicando técnicas de *Data Profiling*.

En este caso se utiliza la técnica de *Pattern Finder* provista por la herramienta Data Cleaner, para verificar el grado de cumplimiento del requerimiento R2 del caso de estudio (i.e. el email del ciudadano debe tener un formato válido). La Figura 15.4 presenta el resultado de aplicar esta técnica sobre el atributo email de la tabla ciudadanos.

Figura 15.4: Resultado de Aplicación de Técnica *Pattern Finder*

La aplicación de esta técnica permite comprobar que el grado de cumplimiento de este requerimiento es total.

La Tabla 15.21 y la Tabla 15.20 presentan cómo se registran estos resultados.

Tabla 15.20: Resultados de *Pattern Finder* sobre Colección Base de Datos SRC IM

Técnica	*Pattern Finder*
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Tabla ciudadanos, atributo email	aaaaaaaaaa@aaaaaaaa.aaa

Tabla 15.21: Aplicar Técnicas *Data Profiling* para Requerimientos de Calidad de Datos

Resultados Técnicas	**Requerimientos**		
	R1 (documento)	**R2 (email)**	**.....**
***Pattern Finder* (1)**		total	

15.3.4. Detectar Nuevos Problemas de Calidad de Datos

En esta actividad se busca detectar nuevos problemas de calidad de datos utilizando técnicas de *Data Profiling*. En particular, para el caso de estudio se utiliza las técnicas provistas por la herramienta DataCleaner y se hace foco en los datos de los reclamos.

En la Figura 15.5 se ve la estructura de datos y algunos datos de ejemplo de los reclamos.

ID_RECLAMO	MOTIVO	FECHA_INGR	ESTADO	FECHA_DESDE	AREA	CATEGORIA	SUBCATEGORIA	COORD_X	COORD_Y
2816	Valentín Alvarez entr	30/12/2016	Finalizado	30/12/2016	Calles y veredas	Viales	Bache	570526.709	6147877.70
2818	Basural fuera conten	1/1/2018	Ingresado	1/1/2018	Limpieza	Estado de los co	Contenedor roto	579076.700	6142999.12
2819	A las viviendas	1/1/2019	Finalizado	22/1/2019	Limpieza	Estado de los co	Solicitar traslado c	578941.188	6137947.96
138019	Cámara principal obs	2/1/2019	Finalizado	2/1/2019	Saneamiento	Conexiones y C	Conexion Obstruid	568710.261	6140716.53
139717	Solicita que se conc	2/1/2017	Finalizado	3/1/2017	Limpieza	Contenedores	No pasó Camión L	579469.875	6141592.64
139718	varios focos	2/1/2018	Finalizado	5/1/2018	Alumbrado	Alumbrado	Problema de Alum	567906.562	6140068.57
139719	UNA PARTE DEL AR	2/1/2019	Finalizado	13/02/2019	Arbolado	Arbolado	Arboles o ramas c	570837.548	6142525.97
140217	Solicita que concurra	2/1/2017	Finalizado	3/1/2017	Limpieza	Contenedores	No paso Camion R	578094.825	6140949.75
158018	Hay un basural fuera	2/1/2018	Ingresado	2/1/2018	Limpieza	Problema de lim	Residuos fuera de	573377.853	6145610.98
158019	Ramas caidas sobre	2/1/2019	Finalizado	11/1/2019	CECOED	Emergencias	CECOED-Arbolad	575347.505	6142439.68
159617	Exp. 2016-3230-98-C	2/1/2017	En Proceso	2/1/2017	Saneamiento	Bocas de Torme	Boca de Tormenta	575505.23	6137254.09
159618	2 cuerpos Veterinaria	2/1/2018	Finalizado	4/1/2018	Limpieza	Problema de lim	Animal muerto de	573596.957	6143019.36
159619	Árbol inclinado a 45	2/1/2019	En Proceso	2/3/2019	Arbolado	Arbolado	Arbol deterorado /	586218.096	6139533.61
160017	Contenedor repleto c	2/1/2017	En Proceso	2/1/2017	Limpieza	Contenedores	No paso Camion R	571108.518	6146611.09
107218	Se encuentran dos b	02/01/18	En Proceso	11/09/13	Calles y veredas	Viales	Bache	570307.493	6146981.7

Figura 15.5: Ejemplos de Datos de Reclamos para el análisis

Por otro lado, en la Figura 15.6 se muestran algunos procesos que se definieron en la herramienta DataCleaner para analizar estos datos. Notar que se utilizan algunos pasos de conversiones porque los datos se importan desde un formato CSV[1] y se cargan como campos de texto.

[1] comma separated values

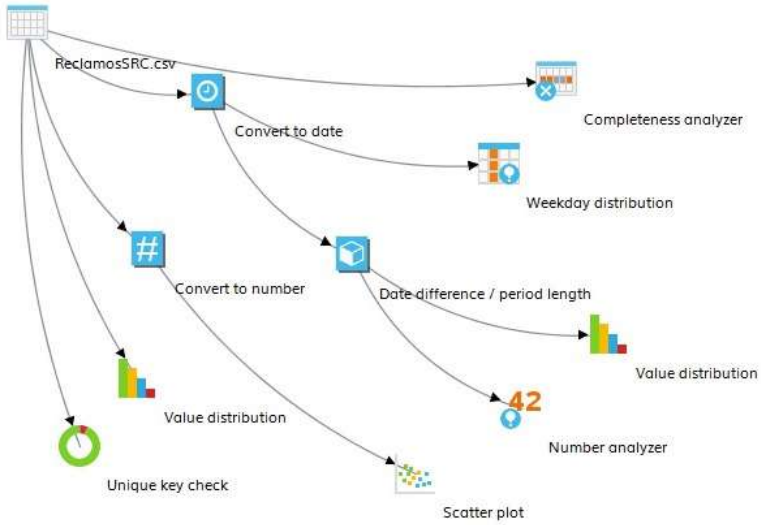


Figura 15.6: Procesos de Análisis de los Datos de Muestra

Ubicación de los Reclamos (DataCleaner)

Como se puede observar en la Figura 15.5, la ubicación de los reclamos se almacena en dos campos numéricos: CoordX y CoordY. La Figura 15.7 presenta el resultado de un análisis *Scatter Plot* utilizando estos campos, el cual permite ver la distribución de la ubicación de los reclamos.

Figura 15.7: Análisis de la Localización de los Reclamos por Ploteo

Aunque DataCleaner no tiene herramientas para trabajar con datos espaciales, utilizando esta forma de graficar se puede observar que los reclamos se distribuyen de una forma similar a la silueta del departamento de Montevideo. En particular, se detecta que un reclamo se encuentra alejado del resto y fuera de los límites de Montevideo.

Tabla 15.22: Resultados de *Scatter Plot* sobre Colección Base de Datos SRC IM

Resultados Técnica	
Técnica	*Scatter Plot*
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Tabla reclamos, atributos coordx-coordy	reclamos fuera de Montevideo

La aplicación de esta técnica permite detectar un nuevo problema de calidad de datos:

- P10: Existen reclamos cuya ubicación está fuera de los límites de Montevideo.

Unicidad de ID de Reclamos (DataCleaner)

Para verificar la unicidad de los identificadores de los reclamos, se utiliza una técnica de validación de unicidad clave (i.e. Unique Key Check). Como se muestra en la Figura 15.8, en este muestreo de datos no hay identificadores repetidos.

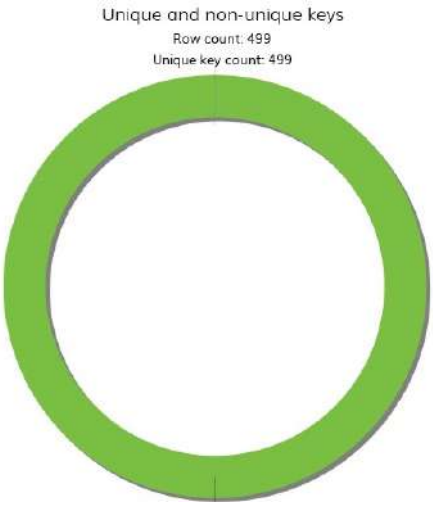
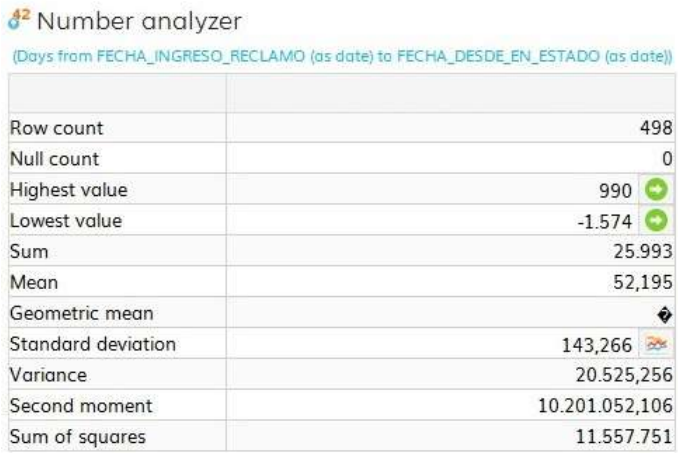


Figura 15.8: Análisis de la Unicidad de los Identificadores de Reclamos

Por lo tanto, con esta técnica no se detecta ningún nuevo problema de calidad de datos.

Correlación entre Fecha de Ingreso del Reclamo y Fecha de Último Estado (DataCleaner)

En este caso se plantea explorar con una técnica de correlación, la relación entre la fecha de ingreso del reclamo y la última fecha de cambio de estado. Para esto se aplica una conversión de datos, cálculo de la diferencia entre las fechas y un análisis numérico de dicha diferencia. Los resultados del análisis numérico se ven en la Figura 15.9.


Number analyzer

(Days from FECHA_INGRESO_RECLAMO (as date) to FECHA_DESDE_EN_ESTADO (as date))

Row count	498
Null count	0
Highest value	990
Lowest value	-1.574
Sum	25.993
Mean	52,195
Geometric mean	
Standard deviation	143,266
Variance	20.525,256
Second moment	10.201.052,106
Sum of squares	11.557.751

Figura 15.9: Diferencia entre Fecha Ingreso Reclamo y Fecha Último Cambio de Estado

Tabla 15.23: Resultados de Análisis Numérico sobre Colección Base de Datos SRC IM

Resultados Técnica	
Técnica	*Análisis Numérico*
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Atributos: Reclamos.fecha, EstadoReclamo.fechaEstado	Valor Mínimo = -1574, valor máximo: 990

La aplicación de esta técnica permite detectar dos nuevos problemas de calidad de datos:

- P11: Existen reclamos que tienen como fecha de último cambio de estado una fecha previa a su ingreso (diferencia entre fechas negativa)
- P12: El valor máximo encontrado para la diferencia entre las fechas (que brinda un indicador del tiempo que demora en resolverse un reclamo) es de 990 días.

Distribución de Valores de Áreas Reclamos

Por último, se utiliza la técnica de distribución de valores para analizar qué valores toman los atributos y cuáles son los más frecuentes. En general no se encuentran problemas de clasificación de los reclamos porque la mayoría de los datos son elegidos por los ciudadanos en base a opciones predefinidas (p. ej. categorías de reclamos).

Sin embargo, como se muestra en la Figura 15.10, se detecta que para el caso de un reclamo aparece el área Arbolado1 en lugar de Arbolado.

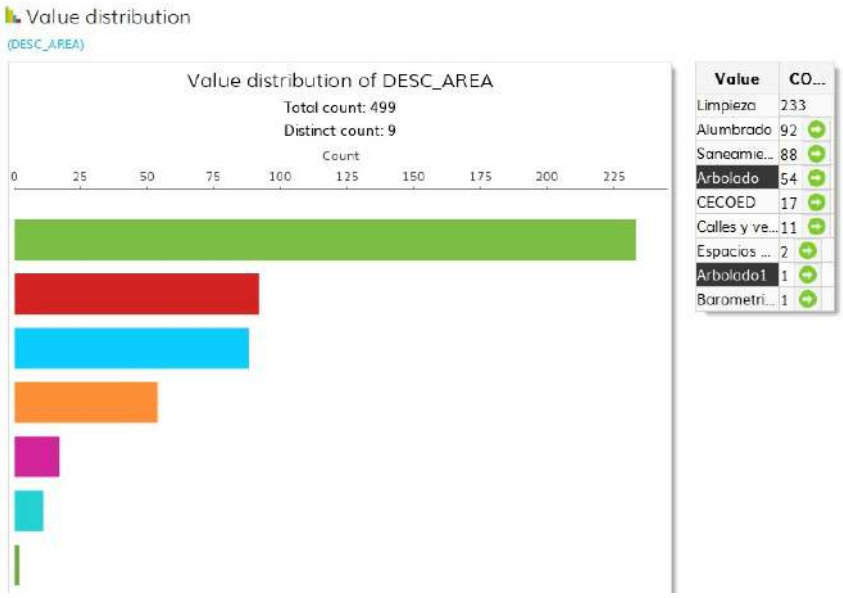


Figura 15.10: Distribución de Valores de Áreas de Atención de Reclamos

Tabla 15.24: Resultados de Distribución de Valores sobre Colección Base de Datos SRC IM

Resultados Técnica	
Técnica	*Distribución de Valores*
Colección de Datos	Base de Datos SRC IM
Fecha	08/11/2019
Configuración	**Resultado**
1. Tablas: Reclamos, SubcategoriasReclamo	SubcategoriasReclamo.nombre = Arbolado1

La aplicación de esta técnica permite detectar un nuevo problema de calidad de datos:

- P13: Existen nombres de categorías que nos son correctos.

Resumen Nuevos Problemas de Calidad de Datos

La Tabla 15.25 presenta un resumen de los nuevos problemas de calidad de datos detectados utilizando técnicas de *Data Profiling*.

Tabla 15.25: Problemas Detectados con Técnicas de *Data Profiling*.

Resultados Técnicas	Problemas			
	P10	P11	P12	P13
Scatter Plot (1)	X			
Análisis Numérico (1)		X	X	
Distribución de Valores (1)				X

15.3.5. Definir Nuevos Requerimientos de Calidad de Datos

En esta actividad se definen nuevos requerimientos de calidad de datos en base a los problemas detectados en la actividad anterior. Los nuevos requerimientos identificados son:

- R13 - La localización del reclamo debe estar dentro de los límites de Montevideo.
- R14 - La fecha de cualquier cambio de estado del reclamo debe ser posterior a la fecha del reclamo.
- R15 - El cambio de estado de un reclamo a «resuelto» debería producirse dentro de las 48 horas posteriores a que el problema haya sido solucionado efectivamente.

La Tabla 15.26 presenta la relación entre los nuevos requerimientos que surgen a partir de los problemas detectados con técnicas de *Data Profiling*.

Tabla 15.26: Relaciones entre Problemas y Requerimientos de Calidad

Problemas	Requerimientos		
	R13	R14	R15
P10	X		
P11		X	
P12			X
P13			

15.4. Definir Estrategia de Gestión de Calidad

Esta sección presenta la etapa Definir Estrategia de Gestión de Calidad en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4), siguiendo las actividades definidas en la Sección 9.3.

15.4.1. Asignar Prioridades a Requerimientos

En este actividad se asignan prioridades a todos los requerimientos identificados en base a los resultados obtenidos en etapas anteriores. La Tabla 15.27 lista los requerimientos identificados en etapas anteriores y la Tabla 15.28 presenta la asignación de prioridades a estos requerimientos.

Tabla 15.27: Requerimientos Identificados

ID	Requerimiento
R1	El número de documento del ciudadano tiene que corresponder a una cédula válida.
R2	El email del ciudadano debe tener un formato de correo electrónico válido.
R3	Los nombres y apellidos de los ciudadanos deben ser los reales.
R4	El email del ciudadano debe existir.
R5	El domicilio del ciudadano debe existir.
R6	El teléfono del ciudadano debe existir
R7	La edad del ciudadano que se registra debe estar entre 10 y 110 años.
R8	El reclamo debe corresponder a un hecho real.
R9	El reclamo no debe estar duplicado.
R10	El estado «rechazado» de un reclamo debería tener una aclaración no vacía.
R11	La localización del reclamo debe ser consistente con su categoría.
R12	La categoría del reclamo debe ser consistente con su fecha.
R13	La localización del reclamo debe estar dentro de los límites de Montevideo.
R14	La fecha de cualquier cambio de estado del reclamo debe ser posterior a la fecha del reclamo.
R15	El cambio de estado de un reclamo a «resuelto» debería producirse dentro de las 48 horas posteriores a que el problema haya sido solucionado efectivamente.

A modo de ejemplo, al requerimiento R1 se le asignó prioridad «alta» dado que es de interés para un gran número de actores y aplica a varios elementos del escenario (p. ej. colecciones, clientes) que son también de interés para varios actores.

Tabla 15.28: Asignación de Prioridades a Requerimientos

Req.	Prioridad	Comentarios
R1	alta	
R2	media	
R3	baja	
R4	alta	
R5	alta	
R6	alta	
R7	baja	
R8	baja	
R9	alta	
R10	alta	
R11	media	
R12	baja	
R13	alta	
R14	alta	
R15	alta	

15.4.2. Describir Estrategia

En este actividad se describe la estrategia de gestión de calidad de datos, en base a la asignación de prioridades a requerimientos. En particular, para el caso de estudio se determina que la estrategia consista en abordar, en primer lugar, los requerimientos con prioridad alta y muy alta, para luego abordar los requerimientos con otras prioridades.

Tabla 15.29: Describir Estrategia

Estrategia de Gestión de Calidad de Datos	
Nombre Estrategia	Estrategia Gestión Calidad para Sistema SRC
Descripción Estrategia	Abordar en primer lugar los requerimientos con prioridad alta y muy alta. Luego abordar los requerimientos con otras prioridades.

De esta forma, la estrategia consiste de dos pasos:

1. Paso 1: se abordan los requerimientos con prioridad alta y muy alta
2. Paso 2: se aborda el resto de los requerimientos

La Tabla 15.30 presenta los requerimientos que se abordan en cada paso de la estrategia.

Tabla 15.30: Pasos de la Estrategia

Requerimiento	Paso 1	Paso 2
R1	1	
R2		X
R3		X
R4	X	
R5	X	
R6	X	
R7		X
R8		X
R9	X	
R10	X	
R11		X
R12		X
R13	X	
R14	X	
R15	X	

15.5. Definir Modelo de Calidad de Datos

Esta sección presenta la etapa Definir Modelo de Calidad de Datos en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 10.3.

15.5.1. Identificar Elementos del Modelo de Referencia

En esta actividad se identifican los elementos del modelo de referencia a utilizar. Para esto, se toma como base los requerimientos incluidos en el primer paso de la estrategia (cf. Sección 9.4.2) y se identifican elementos (i.e. dimensiones, factores y métricas) del modelo de referencia que estén asociados.

La Tabla 15.31 presenta los requerimientos incluidos en el primer paso de la estrategia

La Tabla 15.32 presenta los elementos del modelo de referencia asociados a estos requerimientos.

15.5.2. Definir Elementos Base del Modelo

En esta actividad se definen los elementos del modelo de calidad de datos a construir, tomando como insumo los elementos del modelo de referencia identificados en la actividad anterior.

Tabla 15.31: Requerimientos Identificados

ID	Requerimiento
R1	El número de documento del ciudadano tiene que corresponder a una cédula válida.
R4	El email del ciudadano debe existir.
R5	El domicilio del ciudadano debe existir.
R6	El teléfono del ciudadano debe existir
R9	El reclamo no debe estar duplicado.
R10	El estado «rechazado» de un reclamo debería tener una aclaración no vacía.
R13	La localización del reclamo debe estar dentro de los límites de Montevideo.
R14	La fecha de cualquier cambio de estado del reclamo debe ser posterior a la fecha del reclamo.
R15	El cambio de estado de un reclamo a «resuelto» debería producirse dentro de las 48 horas posteriores a que el problema haya sido solucionado efectivamente.

Tabla 15.32: Problemas / Requerimientos de Calidad y Factores de Calidad asociados

Req.	Dimensión	Factor	Métrica Genérica / Específica
R1	Exactitud	Correctitud Sintáctica	Formato (NumeroDocumento, DNIC)
R4	Exactitud	Correctitud Semántica	CorrectitudSemDebil
R5	Exactitud	Correctitud Semántica	CorrectitudSemDebil
R6	Exactitud	Correctitud Semántica	CorrectitudSemDebil
R9	Unicidad	No-duplicación	EntidadDuplicada
R10	Completitud	Densidad	NoNulo
R13	Consistencia	Integridad Inter-entidad	ReglaEspacial
R14	Consistencia	Integridad Inter-entidad	ReglaIntegridadInterEntidad
R15	Frescura	Actualidad	DesactualizaciónPorFecha

De esta forma, el modelo de calidad de datos para el escenario incluirá las siguientes dimensiones: Exactitud, Unicidad, Completitud, Consistencia y Frescura. Asimismo, incluirá los factores identificados en la Tabla 15.32.

Con respecto a las métricas, a continuación se analiza cómo definir una métrica instanciada para cada uno de los requerimientos a abordar.

El requerimiento R1 implica medir la correctitud sintáctica en documentos de identidad uruguayos. El modelo de referencia posee una métrica específica para dicha situación, por lo que se genera en el modelo una métrica instanciada M1 basada en dicha métrica específica. La métrica M1 se define en la Tabla 15.33.

El requerimiento R4 implica medir la correctitud semántica en direcciones de correo electrónico. El modelo de referencia no posee una métrica específica para dicha situación, pero sí posee una métrica genérica, por lo que se genera en el modelo una métrica instanciada M4 basada en dicha métrica genérica. La métrica M4 se define en la Tabla 15.34.

Tabla 15.33: Definición de la Métrica M1

Métrica M1	
Dimensión:	Exactitud
Factor:	Correctitud Sintáctica
Basada en:	Formato (NumeroDocumento, DNIC)
Nombre:	Formato (numDoc, DNIC)
Semántica:	Indica si el valor del atributo numDoc cumple con el formato de cédula de identidad uruguaya establecido por DNIC, que establece que la cédula tiene siete dígitos seguidos de un octavo dígito verificador que está en función de los otros siete $(d_1 d_2 d_3 d_4 d_5 d_6 d_7 - v, v = f(d_i))$
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	NumeroDocumento=Ciudadano.numDoc Estandar(NumeroDocumento)=«DNIC»

Tabla 15.34: Definición de la Métrica M4

Métrica M4	
Dimensión:	Exactitud
Factor:	Correctitud Semántica
Basada en:	CorrectitudSemDebil
Nombre:	CorrectitudSemDebil(eMail)
Semántica:	Evalúa si una dirección de correo electrónico existe. Se utiliza una función en un lenguaje de alto nivel que permite verificar la existencia a través del envío de un correo electrónico y el análisis de la respuesta.
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	Atr=Ciudadano.eMail Función= existeEmail(eMail: String): Boolean

El requerimiento R5 implica medir la correctitud semántica en domicilios ingresados como texto libre. El modelo de referencia no posee una métrica específica para dicha situación, pero sí posee una métrica genérica, por lo que se genera en el modelo una métrica instanciada M5 basada en dicha métrica genérica. La métrica M5 se define en la Tabla 15.35.

Tabla 15.35: Definición de la Métrica M5

Métrica M5	
Dimensión:	Exactitud
Factor:	Correctitud Semántica
Basada en:	CorrectitudSemDebil
Nombre:	CorrectitudSemDebil(domicilio)
Semántica:	Evalúa si una dirección geográfica existe. Se utiliza una capa geográfica de direcciones de Montevideo como referencial.
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	Atr=Ciudadano.domicilio Referencial= DireccionesMontevideo

El requerimiento R6 implica medir la correctitud semántica en números telefónicos uruguayos. El modelo de referencia no posee una métrica específica para dicha situación, pero sí posee una métrica genérica, por lo que se genera en el modelo una métrica instanciada M6 basada en dicha métrica genérica. La métrica M6 se define en la Tabla 15.36.

Tabla 15.36: Definición de la Métrica M6

Métrica M6	
Dimensión:	Exactitud
Factor:	Correctitud Semántica
Basada en:	CorrectitudSemDebil
Nombre:	CorrectitudSemDebil(telefono)
Semántica:	Evalúa si un número telefónico, existe dentro de un diccionario de números telefónicos de Uruguay. Este diccionario está formado por teléfonos fijos proporcionados por Antel, y teléfonos celulares proporcionados por las operadoras Antel, Movistar y Claro.
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	Atr=Ciudadano.telefono Diccionario=TelefonosFijosAntel $\cup$ TelefonosCelularesAntel $\cup$ TelefonosCelulares-Movistar $\cup$ TelefonosCelularesClaro

El requerimiento R9 implica medir la no-duplicación de los reclamos. El modelo de referencia posee una métrica genérica llamada EntidadDuplicada que se ajusta a este caso, por lo que se genera en el modelo una métrica instanciada M5 basada en dicha métrica genérica. La métrica M9 se define en la Tabla 15.37.

Tabla 15.37: Definición de la Métrica M9

Métrica M9	
Dimensión:	Unicidad
Factor:	No-duplicación
Basada en:	EntidadDuplicada
Nombre:	EntidadDuplicada(Reclamo, [fecha, coordX, coordY, tipo])
Semántica:	Evalúa si una instancia de un reclamo corresponde a la misma situación del mundo real que es reportada en otra instancia. Para determinar que dos instancias corresponden a la misma situación se definen los siguientes criterios: 1. Las fechas de creación de los reclamos debe tener una diferencia máxima de 30 días 2. Los reclamos debe ser del mismo tipo 3. La distancia entre las localizaciones de los reclamos debe ser menor a 50 metros.
Granularidad:	instanciaEntidad
Tipo Resultado:	Boolean
Reglas Correspondencia:	Ent=Reclamo ConjuntoAtributos= [fecha,coordX,coordY,tipo]

El requerimiento R10 implica medir la densidad del atributo observaciones del estado «rechazado» de los reclamos que se encuentran en dicho estado. El modelo de referencia posee una métrica genérica llamada NoNulo que se ajusta a este caso, por lo que se genera en el modelo una métrica instanciada M10 basada en dicha métrica genérica. La métrica M10 se define en la Tabla 15.38.

Tabla 15.38: Definición de la métrica M10

Métrica M10	
Dimensión:	Completitud
Factor:	Densidad
Basada en:	NoNulo
Nombre:	NoNulo (EstadoReclamo.observaciones)
Semántica:	Indica si el atributo «observaciones» de la entidad «EstadoReclamo» tiene un valor nulo o vacío. Se consideran únicamente los estados con tipoEstado=«rechazado».
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	Atr=EstadoReclamo.observaciones DiccionarioValoresNulos(EstadoReclamo.observaciones)={Null, StringVacío}

El requerimiento R13 implica medir la integridad inter-entidad entre la localización del reclamo, dada por el par de atributos (coordX,coordY), y la capa geográfica de departamentos de Uruguay. El modelo de referencia posee una métrica genérica llamada ReglaEspacial que se ajusta a este caso, por lo que se genera en el modelo una métrica instanciada M13 basada en dicha métrica genérica. La métrica M13 se define en la Tabla 15.39.

Tabla 15.39: Definición de la métrica M13

Métrica M13	
Dimensión:	Consistencia
Factor:	Integridad Inter-entidad
Basada en:	ReglaEspacial
Nombre:	ReglaEspacial(Reclamo, Departamento, Whithin)
Semántica:	Indica si el punto donde se ubica el reclamo está dentro del polígono del departamento de Montevideo.
Granularidad:	conjuntoEntidades
Tipo Resultado:	Boolean
Reglas Correspondencia:	Geom1=MakePoint(Reclamo.coordX,Reclamo.coordY) – La función MakePoint genera una geometría de tipo punto para cada par de coordenadas (X, Y) que recibe como parámetro. Geom2=Departamentos(nombre=«Montevideo»).geom – Se utiliza la geometría del departamento de Montevideo obtenida de la capa de polígonos de departamentos. PredicadoEspacial=Within – Geom1.Within(Geom2)

El requerimiento R14 implica medir la integridad inter-entidad entre la fecha de creación del reclamo y las fechas de los estados de ese reclamo. El modelo de referencia posee una métrica genérica llamada ReglaIntegridadInterEntidad que se ajusta a este caso, por lo que se genera en el modelo una métrica instanciada M14 basada en dicha métrica genérica. La métrica M14 se define en la Tabla 15.40.

Tabla 15.40: Definición de la Métrica M14

Métrica M14	
Dimensión:	Consistencia
Factor:	Integridad Inter-entidad
Basada en:	ReglaIntegridadInterEntidad
Nombre:	ReglaIntegridadInterEntidad(Reclamo.fecha, EstadoReclamo.fechaHora)
Semántica:	Indica si el atributo «fecha» de la entidad «Reclamo» es igual o anterior al atributo «fechaHora» de la entidad «EstadoReclamo», medido sobre cada reclamo y todos los estados relacionados con ese reclamo.
Granularidad:	instanciaAtributo
Tipo Resultado:	Boolean
Reglas Correspondencia:	ConjuntoAtributos=(Reclamo.fecha, EstadoReclamo.fechaHora). ExpresionCondicional=Reclamo(id=r).fecha $\leq$ EstadoReclamo(idReclamo=r).fechaHora $\forall r$

El requerimiento R15 implica medir la oportunidad del cambio al estado «resuelto» de un reclamo, comparándolo con la fecha en la que efectivamente se solucionó el problema en la realidad. El modelo de referencia posee una métrica genérica llamada BoolOportunidadPorIntervalo que se ajusta a este caso, por lo que se genera en el modelo una métrica instanciada M15 basada en dicha métrica genérica. La métrica M15 se define en la Tabla 15.41.

Tabla 15.41: Definición de la Métrica M15

Métrica M15	
Dimensión:	Frescura
Factor:	Oportunidad
Basada en:	ReglaEspacial
Nombre:	BoolOportunidadEntPorIntervalo(EstadoReclamo)
Semántica:	Indica si la instancia de la entidad EstadoReclamo de un reclamo dado con tipoEstado=«finalizado», tiene el valor del atributo fechaHora comprendido entre la fecha de resolución del reclamo y las 48 horas siguientes. La fecha de resolución del reclamo se toma del sistema de gestión del área correspondiente.
Granularidad:	instanciaEntidad
Tipo Resultado:	Boolean
Reglas Correspondencia:	Ent=EstadoReclamo(tipoEstado=«finalizado») Atr=EstadoReclamo.fechaHora Referencial=Resolucion(idReclamo) Intervalo=[T_i, T_i+48h], T_i=Resolucion(idReclamo).fechaHora

Por último, para cada una de las métricas definidas se define un posible método para su medición. La Tabla 15.42 presenta una descripción de estos métodos.

Tabla 15.42: Métodos de Calidad de Datos del Modelo

Métrica	Método	Descripción Método
M1	Met1	Se ejecuta una rutina en el lenguaje Java que verifica el largo correcto del número de documento y compara el dígito verificador almacenado con el valor calculado.
M4	Met4	Se ejecuta una rutina en el lenguaje Java que envía un correo electrónico a cada dirección, utilizando el protocolo SMTP, y luego analiza la respuesta para comprobar si la dirección no existe (códigos 551 y 554).
M5	Met5	Se ejecuta una rutina en el lenguaje Java que normaliza y geocodifica la dirección ingresada, utilizando capas de direcciones de Montevideo dadas por nombre de calle y número de puerta. Las direcciones que no pueden ser normalizadas debido a que provienen de un atributo de texto libre, no necesariamente puede considerarse que no existen, por lo que el resultado es una aproximación con cierto margen de error.
M6	Met6	Se ejecuta una consulta SQL buscando los números telefónicos que no se encuetran en ninguna de las tablas referenciales de los números proporcionados por Antel, Movistar y Claro.
M9	Met9	Se ejecuta una consulta SQL sobre la tabla Reclamo, buscando tuplas que cumplan que |r1.fecha-r2.fecha|<30d, r1.tipo=r2.tipo y Distance(MakePoint(r1.coordX,r1.coordY), MakePoint(r2.coordX,r2.coordY))<50m.
M10	Met10	Se ejecuta una consulta SQL haciendo un JOIN entre Reclamo y EstadoReclamo, con la condición de que EstadoReclamo.tipoEstado=«rechazado» y EstadoReclamo.observaciones sea nulo.
M13	Met13	Se ejecuta una consulta SQL haciendo un JOIN ESPACIAL entre Reclamo Y Departamento, con la condición de que Departamento.nombre=Montevideo y Whithin(MakePoint(coordX,coordY), Departamento.geom).
M14	Met14	Se ejecuta una consulta SQL haciendo un JOIN entre Reclamo Y EstadoReclamo, con la condición de que Reclamo.fecha<=EstadoReclamo.fechaHora.
M15	Met15	Se ejecuta una consulta SQL haciendo un JOIN entre Reclamo, EstadoReclamo y Resolucion por idReclamo, con la condición de que EstadoReclamo.tipoEstado=«rechazado» y Resolucion.fecha<=EstadoReclamo.fechaHora<= Resolucion.fecha+48h.

15.5.3. Definir Perfiles de Evaluación

En esta actividad se definen perfiles de evaluación en base a reglas de evaluación para las métricas instanciadas que resultaron de la actividad anterior. En particular, para cada una de las métricas presentadas en la Sección 15.5.2 con tipo de resultado «Boolean», se define una regla de evaluación que tiene como condición que el resultado sea «true».

La Tabla 15.43 presenta la definición de un perfil de evaluación con varias de estas reglas de evaluación.

Tabla 15.43: Perfil de Evaluación

Perfil de Evaluación: Perfil Básico		
Regla	**Métrica**	**Condición**
RegDoc: Cumple Formato Documento	M1	resultado = true
RegEmail: Existe Email	M4	resultado = true
RegDomic: Existe Dirección Geo	M5	resultado = true

15.6. Medir y Evaluar la Calidad de Datos

Esta sección presenta la etapa Medir y Evaluar Calidad de Datos en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 11.3. En particular, se considera que hay definido un único perfil con las reglas definidas en la Tabla 15.43.

15.6.1. Implementar Métodos de Medición

En esta actividad se implementan o adaptan métodos de medición para poder tomar medidas de calidad de datos en el escenario de trabajo dado.

En este caso, se asume que ya se contaba con una implementación base de los métodos requeridos de acuerdo a lo especificado en la Tabla 15.42. Por este motivo, para el caso de estudio esta etapa sólo implica la adaptación de los métodos de medición de forma que puedan acceder a las colecciones de datos para tomar las medidas (p. ej. especificando una cadena de conexión para que puedan acceder una tabla de una base de datos relacional).

15.6.2. Medir Calidad de Datos

En esta actividad, se ejecutan los métodos de medición para tomar las medidas de calidad de datos, de acuerdo a las métricas identificadas, y se almacenan dichas medidas para su posterior evaluación.

La Tabla 15.44 presenta ejemplos de medidas para la métrica «M1: Formato (numDoc, DNIC)» identificada en la Sección 15.5.2.

Tabla 15.44: Medidas Métrica M1: Formato (numDoc, DNIC)

idMedida	idMedición	Fecha	Resultado	idOrganización	idColección	idEntidad	idAtributo	idObj
1	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	numDoc	25
2	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	numDoc	27
3	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	numDoc	28
4	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	numDoc	30

La Tabla 15.45 presenta ejemplos de medidas para la métrica «M4: CorrectitudSemDebil (email)» identificada en la Sección 15.5.2.

Tabla 15.45: Medidas MétricaM4: CorrectitudSemDebil (email)

idMedida	idMedición	Fecha	Resultado	idOrganización	idColección	idEntidad	idAtributo	idObj
10	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	email	25
11	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	email	27
12	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	email	28
13	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	email	30

La Tabla 15.46 presenta ejemplos de medidas para la métrica «M5: CorrectitudSemDebil (domicilio)» identificada en la Sección 15.5.2.

Tabla 15.46: Medidas Métrica M5: CorrectitudSemDebil (domicilio)

idMedida	idMedición	Fecha	Resultado	idOrganización	idColección	idEntidad	idAtributo	idObj
21	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	domicilio	25
22	12	2019-12-06	false	IM	BD SRC IM	Ciudadano	domicilio	27
23	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	domicilio	28
24	12	2019-12-06	true	IM	BD SRC IM	Ciudadano	domicilio	30

15.6.3. Evaluar Calidad de Datos

En esta actividad, se evalúa la calidad de datos en el escenario de trabajo en base a las medidas tomadas así como a las reglas y perfiles de evaluación.

La Tabla 15.47 presenta la evaluación de las distintas medidas tomadas en base a las reglas definidas para el caso de estudio, cuya condición establecía que el resultado de la medida debía tener un valor «true».

Tabla 15.47: Evaluaciones Medidas

Regla	Métrica	idMedida	idMedición	Fecha	Resultado
RegDoc	M1	1	1234	2019-12-02	true
RegDoc	M1	2	1234	2019-12-02	true
RegDoc	M1	3	1234	2019-12-02	false
RegDoc	M1	4	1234	2019-12-02	true
RegEmail	M4	10	1234	2019-12-02	false
RegEmail	M4	11	1234	2019-12-02	true
RegEmail	M4	12	1234	2019-12-02	true
RegEmail	M4	13	1234	2019-12-02	false
RegDomic	M5	21	1234	2019-12-02	false
RegDomic	M5	22	1234	2019-12-02	false
RegDomic	M5	23	1234	2019-12-02	true
RegDomic	M5	24	1234	2019-12-02	true

La Tabla 15.48 presenta la evaluación de las reglas definidas para el caso de estudio, en base a la evaluación de las medidas realizada con dichas reglas.

La Tabla 15.49 presenta la evaluación del perfil definido para el caso de estudio, en base a la evaluación de las reglas asociadas al perfil.

Tabla 15.48: Evaluación Reglas - Caso de Estudio

idMedición	Regla	Fecha	Resultado
1234	RegDoc	2019-12-02	0,75
1234	RegEmail	2019-12-02	0,5
1235	RegDomic	2019-12-04	0,5

Tabla 15.49: Evaluación Perfiles

idMedición	Perfil	Fecha	Resultado
1234	Perfil Básico	2019-12-02	0,58

15.7. Determinar Causas Problemas

Esta sección presenta la etapa Determinar Causas de Problemas de Calidad en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 12.3.

15.7.1. Confirmar y Detectar Problemas de Calidad de Datos

En esta actividad se confirman y detectan problemas de calidad en base a los resultados de la evaluación de la calidad de datos.

Para confirmar los problemas ya identificados para el caso de estudio, se analizan las evaluaciones de las reglas vinculadas, en particular, para los siguientes problemas:

- P1: Correos electrónicos inexistentes
- P2: Domicilios no encontrados

La regla vinculada al problema P1 es RegEmail. Por lo que se puede observar en la Tabla 15.48, la evaluación de esta regla resulta en un valor de 0,55 por lo que se entiende que el problema queda confirmado por esta evaluación.

La regla vinculada al problema P2 es RegDomic. De forma similar, la evaluación de esta regla (cf. Tabla 15.48) tiene un valor de 0,5, por lo que también se entiende que el problema P2 queda confirmado por la misma.

Por otro lado, se puede observar que la evaluación de la regla RegDoc incluida en el Perfil Básico es de 0,75. Esto permite detectar un nuevo problema para el caso de estudio:

- P13: Documentos con formato no válido

La Tabla 15.50 resume este análisis presentando las evaluaciones de qué reglas confirman o detectan los problemas de calidad de datos considerados previamente.

Tabla 15.50: Confirmación y Detección de Problemas de Calidad para Caso de Estudio

Problema	Nombre	Reglas	idMedición
P1	Correos electrónicos inexistentes	RegEmail	1234
P2	Domicilios no encontrados	RegDomic	1234
P13	Documentos con formato no válido	RegDoc	1234

15.7.2. Determinar Causas de Problemas de Calidad

En esta actividad se determinan las causas de los problemas de calidad confirmados en el marco del caso de estudio.

Para esto se analizan las formas en que los datos asociados a los tres problemas considerados (documento de identidad, email y domicilio) fluyen a través de los clientes de datos y son almacenados en las colecciones.

En particular, en la caracterización técnica y de negocio del escenario se puede observar que estos datos son ingresados por los ciudadanos mediante la aplicación móvil y fluyen a través de los servicios web y la aplicación empresarial, que es la que los almacena en la base de datos. Las causas de los problemas de calidad pueden estar asociadas entonces a cualquiera de estos componentes así como a procedimientos generales de la organización.

La Tabla 15.51 presenta las causas que se determinan y confirman en el marco del caso de estudio.

Tabla 15.51: Causas Problemas de Calidad - Caso de Estudio

Problema	Causa	Descripción Causa	Tipo Causa
P1	Falta mecanismo verificación email	Cuando el ciudadano especifica un email no existe un mecanismo que controle que el email existe (p. ej. enviando un correo de confirmación).	Defecto en aplicación
P1	Falta mecanismo actualización email	Si bien el email pudo haber sido especificado en forma correcta por el ciudadano, no existe un procedimiento que asegure que el email se mantenga actualizado (p. ej. para el caso de que el ciudadano no utilice más un email)	Problemas en procedimientos
P2	Falta validación de dirección	Cuando el ciudadano especifica una dirección, no se valida que la misma exista	Defecto en aplicación
P13	Falta validación documento	Cuando el ciudadano especifica un documento, no se valida que su formato sea correcto	Defecto en aplicación

15.8. Definir, Ejecutar y Evaluar Plan Mejora

Esta sección presenta la etapa Definir, Ejecutar y Evaluar Plan de Mejora en el escenario de trabajo planteado en el caso de estudio (cf. Capítulo 4) y siguiendo las actividades definidas en la Sección 13.3.

15.8.1. Definir Plan de Mejora

En esta actividad se define el plan de mejora para el escenario de trabajo en base a un conjunto de acciones de mejora y tomando como base las prioridades de las causas de los problemas de calidad determinados.

La Tabla 15.52 presenta la lista de prioridades para las distintas causas de problemas de calidad de datos determinados para el caso de estudio.

Tabla 15.52: Prioridades - Causas Problemas de Calidad - Caso de Estudio

Problema	Causa	Prioridad
P1	Falta mecanismo verificación email	alta
P1	Falta mecanismo actualización email	alta
P2	Falta validación de dirección	muy alta
P13	Falta validación documento	alta

Por otro lado, la Tabla 15.53 presenta las acciones a tomar en el marco del plan de mejora.

Tabla 15.53: Acciones de Plan de Mejora - Caso de Estudio

Acción	Descripción	Estrategias	Causas
A1	Se debe actualizar la aplicación para que incluya una funcionalidad de verificación de correo electrónico	Corrección defecto	Falta mecanismo verificación email
A2	Se debe establecer un procedimiento (manual o automático) que permita mantener actualizadas las direcciones de correo electrónico	Re-estructura procedimiento	Falta mecanismo actualización email
A3	Se debe actualizar la aplicación para que verifique una dirección al momento que se ingrese	Corrección defecto	Falta validación de dirección
A4	Se debe actualizar la aplicación para que verifique el formato del documento de identidad	Corrección defecto	Falta validación documento

15.8.2. Ejecutar y Evaluar Plan de Mejora

En esta actividad se ejecuta y evalúa el plan de mejora en base a la evaluaciones de los perfiles. A modo de ejemplo, se considera que luego de ejecutar el plan de mejora se obtuvo el resultado que se presenta en la Tabla 15.54 para la evaluación del perfil básico.

Tabla 15.54: Evaluación Perfiles - Plan de Mejora

idMedición	Perfil	Fecha	Resultado
1234	Perfil Básico	2019-12-29	0,80

De esta forma, el nivel de cumplimiento del perfil aumentó en más de un 30 %.

16 Detalle de Dimensiones de Calidad

Este capítulo presenta el detalle de las dimensiones del modelo de calidad de referencia presentado en el Capítulo 3. En particular, se presentan los datos de las métricas definidas en cada dimensión.

16.1. Dimensión Exactitud

Esta sección describe la dimensión Exactitud y sus factores: Correctitud Semántica, Correctitud Sintáctica, Precisión, Exactitud Posicional Absoluta y Exactitud Posicional Relativa. Estos dos últimos factores se utilizan exclusivamente para el el tipo de datos geográfico. La Tabla 16.1 presenta los datos generales de la dimensión.

Tabla 16.1: Dimensión Exactitud

Exactitud	
Definición	Proximidad entre un valor de datos *v* y un valor de datos *v'*, considerado como la representación correcta del fenómeno del mundo real que *v* intenta representar. (Adaptado de [BS16])
Otros nombres	Accuracy (ingl.), correctitud
Factores	Correctitud Semántica Correctitud Sintáctica Precisión Exactitud Posicional Absoluta (TD:«Geo») Exactitud Posicional Relativa (TD:«Geo»)

La Tabla 16.2 presenta los datos generales del factor Correctitud Semántica.

Tabla 16.2: Factor Correctitud Semántica

Correctitud Semántica	
Definición	Proximidad entre el valor *v* de un atributo y su verdadero valor *v'*. (Adaptado de [BS16])
Otros nombres	Semantic Correctness (ingl.)
Métricas Genéricas	***Nombre:*** **CorrectitudSemDebil** ***Semántica:*** Evalúa si una instancia de un atributo, que no forma parte de la identificación de la entidad a la que pertenece, existe dentro de un referencial de valores posibles de ese atributo. ***Granularidad:*** instanciaAtributo ***TipoResultado:*** Boolean ***NombreSugerido:*** CorrectitudSemDebil(Atr) ***PropiedadesConfiguración:*** 1. Id=IdentificaciónAtributo 2. Atributo=Atr 3. Referencial(Id, Atr)
	Nombre: **CorrectitudSemFuerte** ***Semántica:*** Evalúa si una instancia de un atributo, que forma parte de la identificación de la entidad a la que pertenece, existe dentro de un referencial de valores posibles de ese atributo. ***Granularidad:*** instanciaAtributo ***TipoResultado:*** Boolean ***NombreSugerido:*** CorrectitudSemDebil(Atr) ***PropiedadesConfiguración:*** 1. Id=Identificación 2. Atributo=Atr $\in$ Id 3. Referencial(Identificación)
	Nombre: **RatioCorrectitudSemFuerte** ***Semántica:*** Métrica agregada de tipo Ratio basada en CorrectitudSemFuerte. ***Granularidad:*** atributo ***TipoResultado:*** Intervalo real [0, 1]
	Nombre: **RatioCorrectitudSemDébil** ***Semántica:*** Métrica agregada de tipo Ratio basada en CorrectitudSemDébil. ***Granularidad:*** atributo ***TipoResultado:*** Intervalo real [0, 1]

El Ejemplo 50 presenta un problema de Correctitud Semántica.

Ejemplo 50

Con el objetivo de conocer la composición racial de los habitantes y determinar los perfiles demográficos y socio económicos de la población, el Instituto de Estadística releva el dato conocido como Autopercepción de la Ascendencia Racial. Este dato se construye a través de una serie de cinco preguntas de respuesta sí/no en donde el ciudadano contesta si cree tener ascendencia negra, amarilla, blanca, indígena u otra. Con las respuestas se construye un código binario de cinco dígitos que se guarda como un string (p. ej. «10110»). Se encuentran códigos que no cumplen con esa sintaxis, como «1F», «30», «LF», «ESC», «101000101000», etc.

La Tabla 16.3 y la Tabla 16.4 presentan los datos generales del factor Correctitud Sintáctica.

Tabla 16.3: Factor Correctitud Sintáctica

Correctitud Sintáctica	
Definición	Proximidad entre el valor *v* de un atributo y los elementos del dominio de definición de dicho atributo. (Adaptado de [BS16])
Otros nombres	Syntactic Correctness (ingl.)
Métricas Genéricas	***Nombre:*** **Formato** ***Semántica:*** Indica si el valor de un atributo cumple con el formato definido para ese atributo según algún estándar o diccionario. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***NombreSugerido:*** Formato(Atr, Estandar) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. Estandar(Atributo) o Diccionario(Atributo)

Tabla 16.4: Factor Correctitud Sintáctica (continuación)

Correctitud Sintáctica (continuación)	
Métricas Específicas	***Nombre:* Formato(Pais, ISOAlpha3)** ***Semántica:*** Indica si el código de un país está en el formato ISO 3166-1 alpha-3 (3 letras). ***BasadaEn:*** Formato ***PropiedadesConfiguración:*** 1. Atr=Pais 2. Estandar(Pais)=«ISO 3166-1 Alpha-3»
	Nombre:* Formato(Enfermedad, CIE10) [DA:«Salud»]** ***Semántica: Indica si el código de una enfermedad está en el formato CIE-10 (comienza con una letra válida y tiene una cantidad menor o igual a 6 dígitos). ***BasadaEn:*** Formato ***PropiedadesConfiguración:*** 1. Atr=Enfermedad 2. Estandar(Enfermedad)=«CIE-10»
	Nombre:* Formato(NumeroDocumento, DNIC)** ***Semántica: Indica si el valor del atributo numDoc cumple con el formato de cédula de identidad uruguaya establecido por DNIC, que establece que la cédula tiene siete dígitos seguidos de un octavo dígito verificador que está en función de los otros siete ($d_1d_2d_3d_4d_5d_6d_7 - v, v = f(d_i)$). ***BasadaEn:*** Formato ***PropiedadesConfiguración:*** 1. Atr=NumeroDocumento 2. Estandar(NumeroDocumento)=«DNIC»

El Ejemplo 51 presenta un problema de Correctitud Sintáctica.

Ejemplo 51

Con el objetivo de conocer la composición racial de los habitantes y determinar los perfiles demográficos y socio económicos de la población, el Instituto de Estadística releva el dato conocido como Autopercepción de la Ascendencia Racial. Este dato se construye a través de una serie de cinco preguntas de respuesta sí/no en donde el ciudadano contesta si cree tener ascendencia negra, amarilla, blanca, indígena u otra. Con las respuestas se construye un código binario de cinco dígitos que se guarda como un string (p. ej. «10110»). Se encuentran códigos que no cumplen con esa sintaxis, como «1F», «30», «LF», «ESC», «101000101000», etc.

La Tabla 16.5 presenta los datos generales del factor Precisión.

Tabla 16.5: Precisión

Precisión	
Definición	Captura el grado de detalle que posee un dato que lo hace útil para un determinado uso o que permite discriminarlo de otros datos que no son exactamente iguales. (Adaptado de [BS16]) y [ISO08])
Otros nombres	Precision (ingl.)
Métricas Genéricas	***Nombre:*** **Escala** ***Semántica:*** En el caso de valores numéricos, se calcula como $1 - \frac{error}{valor}$n en donde error está dado por el instrumento de medición y valor es el valor del dato. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Intervalo real [0, 1]. ***NombreSugerido:*** Escala(Atr, Error) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. Error=error del instrumento de medición.
	Nombre: **ErrorEstandar** ***Semántica:*** Desviación estándar de un conjunto de datos ***Granularidad:*** Atributo. ***TipoResultado:*** Número real. ***NombreSugerido:*** ErrorEstandar(Atr) ***PropiedadesConfiguración:*** 1. Atributo=Atr

El Ejemplo 52 presenta un problema de Precisión.

Ejemplo 52

En un sistema de trazabilidad de procesos de negocio se registra, entre otros datos, un *timestamp* con la fecha y hora de inicio de un paso de proceso de negocio y otro con la fecha y hora de fin del mismo. Se encuentra que el sistema está guardando esos *timestamps* con el formato «yyyy-MM-dd» en lugar de «yyyy-MM-dd HH:mm:ss». Por este motivo no es posible calcular la duración exacta de un proceso, particularmente en aquellos casos de procesos de corta duración que se completan en horas o minutos, y no en días.

La Tabla 16.6 presenta los datos generales del factor Exactitud Posicional Absoluta.

Tabla 16.6: Factor Exactitud Posicional Absoluta [TD:«Geo»]

Exactitud Posicional Absoluta [TD:«Geo»]	
Definición	Proximidad de los valores reportados de las coordenadas a los valores verdaderos o aceptados como tales [ISO13]
Otros nombres	Absolute positional accuracy (ingl.), Exactitud posicional externa.
Métricas Genéricas	***Nombre:*** **ÍndiceErroresPosicionalesPorUmbral** ***Semántica:*** Ratio entre el número de incertidumbres posicionales superiores a un umbral dado para un conjunto de posiciones, sobre número total de las posiciones medidas (Tabla D.33 de [ISO13]). Esta métrica suele ser evaluada a través de muestreos. ***Granularidad:*** Entidad ***TipoResultado:*** Intervalo real [0, 1].
	Nombre: **ValorMedioIncertidumbrePosicional** ***Semántica:*** Distancia entre la posición medida y la que se considera como verdadera (Tabla D.29 de [ISO13]). Esta métrica suele ser evaluada a través de muestreos. ***Granularidad:*** instanciaEntidad ***TipoResultado:*** Número Real (en las unidades de coordenadas).

La Tabla 16.7 presenta los datos generales del factor Exactitud Posicional Relativa.

Tabla 16.7: Factor Exactitud Posicional Relativa [TD:«Geo»]

Exactitud Posicional Relativa (TD:«Geo»)	
Definición	Proximidad de las posiciones relativas de los objetos geográficos de un conjunto de datos a sus respectivas posiciones relativas verdaderas o aceptadas como tales [ISO13]
Otros nombres	Relative positional accuracy (ingl.), Exactitud posicional interna.
Métricas Genéricas	***Nombre:*** **ErrorHorizontalRelativo** ***Semántica:*** Evaluación de los errores aleatorios en la posición horizontal de una entidad geográfica en relación a otra de la misma capa geográfica (Tabla D.55 de [ISO13]). ***Granularidad:*** instanciaEntidad ***TipoResultado:*** Número Real (en las unidades de coordenadas).

Es importante mencionar que para las métricas de exactitud posicional, es usual recurrir a métodos que sean por lo menos tres veces más precisos que aquello que se desea controlar. Para esto se puede recurrir a: relevamientos directos en campo, imágenes satelitales y aéreas, u otros conjuntos de datos. Es importante que los controles se realicen considerando el universo de discurso (vista del mundo real o hipotético que incluye todo aquello que es de interés [ISO13]) y no la realidad en toda su expresión.

16.2. Dimensión Consistencia

Esta sección describe la dimensión Consistencia y sus factores: Integridad Intra-entidad, Integridad Inter-entidad, Integridad de Dominio y Consistencia Topológica. La Tabla 16.8 presenta los datos generales de la dimensión.

Tabla 16.8: Dimensión Consistencia

Consistencia	
Definición	Captura la violación de las reglas semánticas definidas sobre un conjunto de entidades de negocio o de sus atributos. En un modelo relacional, las restricciones de integridad son un ejemplo de tales reglas semánticas. (Adaptado de [BS16])
Otros nombres	Consistency (ingl.), cohesión, coherencia
Factores	Integridad Inter-entidad Integridad Intra-entidad Integridad de Dominio Consistencia Topológica (TD:«Geo»)

La Tabla 16.9 y la Tabla 16.10 presentan los datos generales del factor Integridad Inter-entidad.

Tabla 16.9: Factor Integridad Inter-entidad

Integridad Inter-entidad	
Definición	Captura la satisfacción de reglas entre atributos de diferentes entidades de negocio. (Adaptado de [BS16])
Otros nombres	Referential Integrity (ingl.), Integridad Inter-relacion
Métricas Genéricas	***Nombre:*** **ReglaIntegridadInterEntidad** ***Semántica:*** Regla de inclusión (clave foránea) o expresión condicional entre atributos de diferentes entidades. ***Granularidad:*** conjuntoEntidades. ***TipoResultado:*** Boolean. ***NombreSugerido:*** ReglaIntegridadInterEntidad(AtrJ,...AtrK) ***PropiedadesConfiguración:*** 1. ConjuntoAtributos={EntX.AtrJ,...,EntY.AtrK } 2. ExpresiónCondicional(ConjuntoAtributos)
	Nombre: **ReglaEspacial [TD:«Geo»]** ***Semántica:*** Expresión condicional que verifica la ocurrencia de determinada relación espacial topológica entre dos atributos geométricos de diferentes entidades. ***Granularidad:*** conjuntoEntidades. ***TipoResultado:*** Boolean. ***NombreSugerido:*** ReglaEspacial(Ent1, Ent2, PredicadoEspacial) ***PropiedadesConfiguración:*** 1. Geom1: Atributo geométrico de la entidad Ent1 2. Geom2: Atributo geométrico de la entidad Ent2 3. PredicadoEspacial: Operador espacial de tipo Boolean (Intersects, Whithin, Contains, Touches, Crosses, Covers, CoveredBy, Overlaps, Equals, Disjoint)

Tabla 16.10: Factor Integridad Inter-entidad (continuación)

Integridad Inter-entidad (continuación)	
Métricas Específicas	***Nombre:*** **ReglaIntegridadInterEntidad(Sexo, Enfermedad) [DA:«Salud»]** ***Semántica:*** Indica si una enfermedad especificada en una entidad de negocio relacionada a una persona (p. ej. historia clínica, certificado de defunción) es compatible con el sexo de esa persona. ***BasadaEn:*** ReglaIntegridadInterEntidad ***PropiedadesConfiguración:*** 1. ConjuntoAtributos={Sexo, Enfermedad } 2. ExpresionSexoEnfermedad(ConjuntoAtributos)

El Ejemplo 53 presenta un problema de Integridad Inter-entidad que también se considera de consistencia lógica según [ISO13].

Ejemplo 53

En un determinado conjunto de datos geográficos, existe una capa de puntos con datos de Aeropuertos y otra de polígonos con datos de Lagos. En la capa de Aeropuertos se encuentra un aeropuerto cuya geometría está contenida en la geometría de un lago de la otra capa, lo que no es posible que suceda en la realidad.

La Tabla 16.11 y la Tabla 16.12 presentan los datos generales del factor Integridad Intra-entidad.

Tabla 16.11: Factor Integridad Intra-entidad

Integridad Intra-entidad	
Definición	Captura la satisfacción de reglas entre atributos de una misma entidad. (Adaptado de [BS16])
Otros nombres	Relation Integrity (ingl.)
Métricas Genéricas	***Nombre:*** **ReglaIntegridadIntraEntidad** ***Semántica:*** Reglas de dependencia de clave y unicidad de atributos, de dependencias funcionales o de dependencias de atributos. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***NombreSugerido:*** ReglaIntegridadIntra(AtrJ,...AtrK) ***PropiedadesConfiguración:*** 1. ConjuntoAtributos={AtrJ,...,AtrK } 2. ExpresiónCondicional(ConjuntoAtributos)
	Nombre: **RatioIntegridadIntraEntidad** ***Semántica:*** Porcentaje de datos que satisfacen una métrica de Integridad Intra-entidad. ***Granularidad:*** atributo. ***TipoResultado:*** Intervalo real [0.0, 1.0] ***NombreSugerido:*** RatioIntegridadIntra(AtrJ,...AtrK) ***PropiedadesConfiguración:*** 1. ConjuntoAtributos={Atr1,...,AtrN } 2. ExpresiónCondicional(ConjuntoAtributos)

Tabla 16.12: Factor Integridad Intra-entidad (continuación)

Integridad Intra-entidad (continuación)	
Métricas Específicas	***Nombre:*** **ReglaIntegridadIntraEntidad(Sexo,Enfermedad) [DA:«Salud»].** ***Semántica:*** Indica si una enfermedad de una persona es compatible con el sexo de esa persona. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***BasadaEn:*** ReglaIntegridadIntraEntidad ***PropiedadesConfiguración:*** 1. ConjuntoAtributos={Sexo, Enfermedad } 2. ExpresionSexoEnfermedad(ConjuntoAtributos)

El Ejemplo 54 y el Ejemplo 55 presentan problemas de Integridad Intra-entidad.

Ejemplo 54

En una tabla llamada Direcciones donde se guardan datos alfanuméricos de direcciones de Uruguay, se encuentra una tupla con atributos nombreCalle=«Coronel Alegre» y nombre-Departamento=«Rocha». El problema consiste en que no existe la calle Coronel Alegre en el departamento de Rocha, por lo tanto hay un error en alguno de los dos atributos.

Ejemplo 55

En una tabla llamada Consultas de la historia clínica de un paciente, se encuentra una tupla con un paciente con los atributos sexo=1 (masculino) y diagnostico=«N80.0», que corresponde al código CIE-10 de la «Endometriosis de útero», que no puede ocurrir en un paciente de sexo masculino.

La Tabla 16.13 y la Tabla 16.14 presentan los datos generales del factor Integridad de Dominio.

Tabla 16.13: Factor Integridad de Dominio

Integridad de Dominio	
Definición	Captura la satisfacción de reglas sobre los valores posibles que puede tomar un atributo. (Adaptado de [BS16])
Otros nombres	Domain Integrity (ingl.)
Métricas Genéricas	***Nombre:*** **ValoresPosiblesPorExtensión** ***Semántica:*** Indica si el valor de un atributo se encuentra dentro de un dominio definido por extensión. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***NombreSugerido:*** ValoresPosiblesPE(Atr) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. DominioPorExtension(Atributo)
	Nombre: **ValoresPosiblesPorComprensión** ***Semántica:*** Indica si el valor de un atributo se encuentra dentro de un dominio definido por comprensión, el cual puede estar dado por una propiedad que cumplen los elementos de ese dominio o por el tipo de dato conocido de ese dominio. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean ***NombreSugerido:*** ValoresPosiblesPC(Atr) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. DominioPorComprensión(Atributo)

Tabla 16.14: Factor Integridad de Dominio (continuación)

Integridad de Dominio (continuación)	
Métricas Específicas	***Nombre:*** **ValoresPosiblesPE(Sexo, AGESIC)** ***Semántica:*** Indica si el valor de un atributo Sexo se encuentra dentro del conjunto de valores definidos en el Vocabulario de Persona de AGESIC ([AGE18]) . ***BasadaEn:*** ValoresPosiblesPorExtensión ***PropiedadesConfiguración:*** 1. Atributo=Sexo 2. DominioPorExtension(Atributo)={(1-Masculino, 2-Femenino, 3-Desconocido, 4-Indeterminado, 9-NoAplica)}

El Ejemplo 56 y el Ejemplo 57 presentan problemas de integridad de dominio. El Ejemplo 58 describe un problema de integridad de dominio, que también se considera de consistencia de dominio según [ISO13].

Ejemplo 56

En una tabla CertificadosNacidosVivos en donde se guardan datos de nacimientos, una tupla tiene un atributo semanasGestacion=400. Este valor está fuera del rango establecido entre 26 y 52 semanas, ya que se considera que duraciones menores o mayores son altamente improbables para nacidos vivos.

Ejemplo 57

En una tabla de Personas, que utiliza el Vocabulario de Persona de AGESIC ([AGE18]), una tupla tiene el atributo tipoDocumento=69999, que no existe dentro de los códigos definidos actualmente por UNAOID (el código mayor es el 69096) para los tipos de documento de identificación de una persona (cédula de identidad, pasaporte, pasaporte diplomático, etc).

Ejemplo 58

En una capa geográfica de polígonos que corresponden a Planes de Uso de Suelo (PDU), algunos PDU tienen el atributo cultivoInvierno=«Soja», pero los valores permitidos para cultivos de invierno son: Barbecho, Cebada, Cereales de invierno, Colza, Cultivo de cobertura, Pastura Consociada, Pastura no consociada, Pasturas y Trigo.

La Tabla 16.15 presenta los datos generales del factor Consistencia Topológica.

Tabla 16.15: Factor Consistencia Topológica

Consistencia Topológica [TD:«Geo»]	
Definición	Corrección de las características topológicas codificadas explícitamente. Las características topológicas de un conjunto de datos describen las relaciones geométricas entre los ítems del conjunto de datos que no son alteradas por transformaciones elásticas (rubber-sheet transformations) [ISO13]
Otros nombres	Topological Consistency (ingl.)
Métricas Genéricas	***Nombre:*** **ÍndiceFallosConexiónNodosEnlace** ***Semántica:*** Porcentaje de fallos en las conexiones de nodos de enlace del total de conexiones de nodos de enlace, medido sobre la geometría de una entidad geográfica (Adaptado de [ISO13] [IDE18]). ***Granularidad:*** atributo. ***TipoResultado:*** Intervalo real [0.0, 1.0].

El Ejemplo 59 describe un problema de consistencia topológica.

Ejemplo 59

En una capa geográfica de líneas que corresponden a ejes de calles de Montevideo, el eje de Ejido hacia el norte de Isla de Flores tiene un nodo de enlace con el eje de Isla de Flores, y el eje hacia el sur tiene otro nodo de enlace, cuando los dos ejes de Ejido deberían llegar al mismo nodo de enlace sobre la calle Isla de Flores.

16.3. Dimensión Completitud

Esta sección describe la dimensión Completitud y sus factores: Cobertura, Densidad y Comisión. La Tabla 16.16 presenta los datos generales de la dimensión.

Tabla 16.16: Dimensión Completitud

Completitud	
Definición	Captura la medida en que los datos son de la amplitud, profundidad y alcance suficientes para una determinada tarea. (Adaptado de [BS16])
Otros nombres	Completeness (ingl.)
Factores	Cobertura Densidad Comisión (TD:«Geo»)

La Tabla 16.17 presenta los datos generales del factor Cobertura.

Tabla 16.17: Factor Cobertura

Cobertura	
Definición	Captura la proporción entre la cantidad de entidades existentes en una determinada colección de datos, y el total de entidades que deberían existir en dicha colección. La cobertura varía si se utiliza la Asunción de Mundo Cerrado, según la cual una colección de datos debería contener todas las entidades de un tipo, o si se utiliza la Asunción de Mundo Abierto, según la cual una colección de datos puede ser una representación parcial de las entidades del mundo real. (Adaptado de [BS16])
Otros nombres	Coverage (ingl.)
Métricas Genéricas	***Nombre:*** **RatioCobertura** ***Semántica:*** Proporción entre la cantidad de instancias de una entidad y el número total de instancias de un referencial de esa entidad. ***Granularidad:*** entidad. ***TipoResultado:*** Intervalo real [0.0, 1.0]. ***NombreSugerido:*** RatioCobertura(Entidad, Referencial) ***PropiedadesConfiguración:*** 1. Entidad=Ent 2. Referencial(Entidad)

El Ejemplo 60 describe un problema de cobertura.

Ejemplo 60

El Ministerio del Interior posee una tabla con los hurtos registrados en 2018. Los números anuales de 2018 se consideran útiles para realizar comparaciones con años anteriores (Asunción de Mundo Cerrado), pero existe información de que muchos hurtos no son denunciados por lo que no figuran en esa tabla (Asunción de Mundo Abierto).

La Tabla 16.18 y la Tabla 16.19 presentan los datos generales del factor Densidad.

Tabla 16.18: Factor Densidad

Densidad	
Definición	Captura la proporción entre la cantidad de instancias de atributo con valores no nulos y el total de instancias de dicho atributo (Adaptado de [BS16]). Un valor nulo de una instancia de atributo *A* de una entidad *E* puede interpretarse de varias maneras: 1. *E* no posee *A* 2. se desconoce si *E* posee *A* o no 3. *E* posee *A* pero se desconoce su valor
Otros nombres	Density (ingl.)
Métricas Genéricas	***Nombre:*** **NoNulo** ***Semántica:*** Indica si una instancia de atributo tiene un valor no nulo. Puede ser necesario especificar un diccionario con todos los valores del atributo que se consideran nulos o vacíos, cuando existe más de una posibilidad. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***NombreSugerido:*** NoNulo(Atr) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. DiccionarioValoresNulos(Atributo)
	Nombre: **DensidadPonderada** ***Semántica:*** Aplica un cálculo sobre algunos atributos de una instancia de entidad, evaluado para cada uno si es nulo o no (como en la métrica NoNulo) pero multiplicando el resultado de cada atributo por un coeficiente entre 0 y 1 cuya suma sea igual a 1. A mayor gravedad de tener un nulo en un atributo, más cercano a 0 será el coeficiente para ese atributo. ***Granularidad:*** instanciaEntidad. ***TipoResultado:*** Intervalo real [0.0, 1.0]. ***NombreSugerido:*** DensidadPonderada(Atr1,...,AtrN) ***PropiedadesConfiguración:*** 1. ConjuntoAtributos=(Atr1,...,AtrN) 2. ConjuntoCoeficientes=(C1,...,CN) 3. DiccionarioValoresNulos(ConjuntoAtributos)

La Tabla 16.20 presenta los datos generales del factor Comisión.

Tabla 16.19: Factor Densidad (continuación)

Densidad (continuación)	
Métricas Genéricas (cont.)	***Nombre:*** **RatioNoNulos** ***Semántica:*** Métrica agregada de tipo Ratio basada en NoNulo. ***Granularidad:*** atributo. ***TipoResultado:*** Intervalo real [0, 1].
	Nombre: **RatioDensidadPonderada** ***Semántica:*** Métrica agregada de tipo Ratio basada en DensidadPonderada. ***Granularidad:*** entidad. ***TipoResultado:*** Intervalo real [0, 1]

Tabla 16.20: Factor Comisión (TD:«Geo»)

Comisión [TD:«Geo»]	
Definición	Datos excedentes presentes en un conjunto de datos [ISO13]
Otros nombres	Commission (ingl.)
Métricas Genéricas	***Nombre:*** **ÍtemExcedente** ***Semántica:*** Indica si una instancia está incorrectamente presente en el conjunto de instancias de una entidad (Tabla D.1 de [ISO13]). ***Granularidad:*** instanciaEntidad ***TipoResultado:*** Boolean. ***PropiedadesConfiguración:***
	Nombre: **ÍndiceItemsExcedentes** ***Semántica:*** Número de entidades excedentes en el conjunto total de instancias de una determinada entidad o muestra de datos de esa entidad, en relación al número del total que deberían haber estado presentes (Tabla D.3 de [ISO13]). ***Granularidad:*** entidad ***TipoResultado:*** Intervalo real [0.0, 1.0]. ***PropiedadesConfiguración:***

16.4. Dimensión Unicidad

Esta sección describe la dimensión *Unicidad* y sus factores: *No-duplicación* y *No-contradicción*. La Tabla 16.21 presenta los datos generales de la dimensión.

Tabla 16.21: Dimensión Unicidad

Unicidad	
Definición	Captura el grado en el que un dato del mundo real es representado en forma única. (Adaptado de [BS16])
Otros nombres	Uniqueness (ingl.). En [BS16] se habla también de *redundancia* para el caso de *linked data*. La norma [ISO13] la considera como parte del factor Compleción y lo establece como una medida para este factor.
Factores	No-duplicación No-contradicción

La Tabla 16.22 y la Tabla 16.23 presentan los datos generales del factor *no-duplicación*.

Tabla 16.22: Factor No-duplicación

No-duplicación	
Definición	Captura el grado de duplicación (o repetición) de un mismo dato.
Otros nombres	Duplication-free (ingl.)
Métricas Genéricas	***Nombre:*** **AtributoDuplicado** ***Semántica:*** Indica si una instancia de atributo tiene el mismo valor que otra instancia del mismo atributo. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***NombreSugerido:*** AtributoDuplicado(Atr) ***PropiedadesConfiguración:*** 1. Atributo=Atr
	Nombre: **ConjuntoAtributosDuplicado** ***Semántica:*** Indica si las instancias de un conjunto de atributos de una instancia de entidad, tienen el mismo valor en otra instancia de entidad. ***Granularidad:*** instanciaEntidad. ***TipoResultado:*** Boolean. ***NombreSugerido:*** ConjuntoAtributosDuplicado(Atr1,...,AtrN) ***PropiedadesConfiguración:*** 1. ConjuntoAtributos=(Atr1,...,AtrN)
	Nombre: **EntidadDuplicada** ***Semántica:*** Indica si existe, para una instancia de entidad, al menos otra instancia más que representa el mismo objeto del mundo real, con los mismos datos o algún dato faltante. Se debe especificar un conjunto de atributos que permita identificar unívocamente una instancia de entidad. Los demás atributos de la entidad que no pertenezcan a dicho conjunto deben tener los mismos valores en las dos instancias, o ser nulos en alguna de ellas, para que se consideren duplicados exactos. ***Granularidad:*** instanciaEntidad. ***TipoResultado:*** Boolean. ***NombreSugerido:*** EntidadDuplicada(Ent, ConjuntoAtributos) ***PropiedadesConfiguración:*** 1. Entidad=Ent 2. ConjuntoClave=(Atr1,...,AtrN)

Tabla 16.23: Factor No-duplicación (continuación)

No-duplicación (continuación)	
Métricas Genéricas	***Nombre:*** **RatioAtributoDuplicado** ***Semántica:*** Métrica agregada de tipo Ratio basada en AtributoDuplicado. ***Granularidad:*** atributo. ***TipoResultado:*** Intervalo real [0, 1].
	Nombre: **RatioConjuntoAtributosDuplicado** ***Semántica:*** Métrica agregada de tipo Ratio basada en ConjuntoAtributosDuplicado. ***Granularidad:*** entidad. ***TipoResultado:*** Intervalo real [0, 1]
	Nombre: **RatioEntidadesDuplicadas** ***Semántica:*** Métrica agregada de tipo Ratio basada en EntidadDuplicada. ***Granularidad:*** entidad. ***TipoResultado:*** Intervalo real [0, 1]

El Ejemplo 61 y el Ejemplo 62 presentan problemas de duplicación.

Ejemplo 61

En un sistema de Historia Clínica Digital, se intercambian mensajes HL7 ADT (Admit Discharge Transfer) que contienen información demográfica del paciente. En cada mensaje, además del identificador principal del paciente, opcionalmente se puede incluir un identificador alternativo o *Alternate Patient ID*. Con el fin de tener datos estadísticos de la cantidad de pacientes que son atendidos por día, otro sistema guarda una tabla diaria con un único registro por paciente atendido cada día. Se detecta que en esa tabla, aparece más de un registro con el mismo *Alternate Patient ID*.

Ejemplo 62

En un sistema de Historia Clínica Digital, se está estudiando la implementación de un Enterprise Master Patient Index (EMPI), para poder identificar datos heterogéneos que corresponden al mismo paciente en distintos sistemas y así mantener un índice único de pacientes. Se quiere dimensionar el problema y justificar la inversión, realizando un estudio de datos duplicados, buscando la repetición exacta de diferentes combinaciones de atributos: a - (primer nombre, primer apellido, fecha de nacimiento), b - (primer nombre, segundo nombre, primer apellido, segundo apellido, fecha de nacimiento), c - (primer nombre, primer apellido, fecha de nacimiento, sexo). Los resultados muestran la existencia de duplicados: a - 20 %, b - 16 %, c - 14 %.

La Tabla 16.24 presenta los datos generales del factor No-contradicción.

Tabla 16.24: Factor No-contradicción

No-contradicción	
Definición	Captura el grado de duplicación (o repetición) de una misma instancia de entidad del mundo real que es representada con datos contradictorios.
Otros nombres	Contradiction-free (ingl.)
Métricas Genéricas	***Nombre:*** **EntidadContradictoria** ***Semántica:*** Indica si existe, para una instancia de entidad, al menos otra instancia más que representa el mismo objeto del mundo real, con alguna contradicción entre sus datos. Dado que las dos instancias de entidad pueden tener distinta clave, se debe especificar una función de similitud que compare un conjunto de atributos en ambas para detectar si es la misma entidad (técnica de resolución de entidades"). La función de similitud puede utilizar la distancia de Levenshtein, la correspondencia de trigramas y algoritmos fonéticos como Soundex, Metaphone, etc. (Capítulo 8 de [BS16]) para detectar las entidades iguales dentro de un cierto umbral. En el caso que las dos entidades tengan los mismos valores en ese conjunto de atributos, sólo se considerarán contradictorias si tienen valores diferentes en los demás atributos, ya que en caso contrario serían entidades duplicadas y no contradictorias. ***Granularidad:*** instanciaEntidad. ***TipoResultado:*** Intervalo real [0.0, 1.0]. ***NombreSugerido:*** EntidadContradictoria(Ent, ConjuntoAtributos, FuncionSimilitud) ***PropiedadesConfiguración:*** 1. Entidad=Ent 2. ConjuntoAtributos=(Atr1,...,AtrN) 3. FuncionSimilitud=F(ConjuntoAtributos, Umbrales)
	Nombre: **RatioEntidadContradictoria** ***Semántica:*** Métrica agregada de tipo Ratio basada en EntidadContradictoria ***Granularidad:*** entidad. ***TipoResultado:*** Intervalo real [0.0, 1.0]

El Ejemplo 63 presenta problemas de contradicción.

Ejemplo 63

En una empresa de telecomunicaciones se está trabajando en la unificación de sus sistemas de telefonía fija y móvil, para lo que se generó una única tabla de clientes en base a su documento de identidad. Existen dudas de la confiabilidad de ese dato en los sistemas de origen, ya que se encontraron entidades duplicadas con contradicciones como las siguientes:

1. c1=(documento:2837451-6, nombre:«Mariana», apellido:«Rinaldi», direccion:«Comandante Braga 2060», telefono: 099123321, profesion:«estudiante»)
2. c2=(documento:2837451-3, nombre«María», apellido:«Rinaldi», direccion:«Comandante Braga 2060», telefono: 098116622, profesion:«bailarina»)

16.5. Dimensión Frescura

Esta sección describe la dimensión *Frescura* y sus factores: *Actualidad* y *Oportunidad*. La Tabla 16.25 presenta los datos generales de la dimensión.

Tabla 16.25: Dimensión Frescura

Frescura	
Definición	Captura la rapidez con la que los cambios en el mundo real son reflejados en la actualización de los datos. La frescura es un tipo de exactitud no-estructural dependiente de la variable tiempo. (Adaptado de [BS16])
Otros nombres	Freshness (ingl.), exactitud temporal
Factores	Actualidad Oportunidad

La Tabla 16.26 y la Tabla 16.27 presentan los datos generales del factor *Actualidad*.

El Ejemplo 64 presenta problemas de actualidad.

Ejemplo 64

En un sistema donde se registran periódicamente las consultas médicas de niños, se encuentra que la estatura de algunos niños no cambió en sucesivas consultas mensuales, por lo que se determina que esos valores están desactualizados.

La Tabla 16.28 presenta los datos generales del factor *Oportunidad*.

Tabla 16.26: Factor Actualidad

Actualidad	
Definición	Captura el tiempo de demora entre un cambio en el mundo real y la correspondiente actualización de los datos.
Otros nombres	Currency (ingl.)
Métricas Genéricas	***Nombre:*** **DesactualizaciónPorFecha** ***Semántica:*** Sea t_a la fecha de acceso a un dato (por defecto es la fecha actual), t_w la fecha del último cambio del dato en el mundo real (o en otra colección de datos de referencia) y t_u la fecha de la última actualización del dato en la colección objetivo. Si $t_u \geqslant t_w$, la métrica devuelve 0. Si $t_u < t_w$, la métrica devuelve la diferencia t_u - t_w. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Real. ***NombreSugerido:*** DesactualizaciónPorFecha(Atr) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. T_w=Fecha del último cambio del dato en el mundo real o en otra colección de datos de referencia. Si se desconoce, se puede calcular en base a la frecuencia de cambio y la fecha de la última actualización del dato. 3. T_u=Fecha de la última actualización del dato. 4. FrecuenciaCambio=Frecuencia con la que cambia el dato (ej. una vez por año). Puede ser estimada en base a la frecuencia de cambio promedio.
	Nombre: **DesactualizaciónPorCambios** ***Semántica:*** Indica la cantidad de cambios sufridos por un dato en el mundo real (o en otra colección de datos de referencia) luego de la última actualización del dato en la colección objetivo. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Entero. ***NombreSugerido:*** DesactualizaciónPorCambios(Atr). ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. T_u=Fecha de la última actualización del dato. 3. FrecuenciaCambio=Frecuencia con la que cambia el dato (ej. una vez por año). Puede ser estimada en base a la frecuencia de cambio promedio.
	Nombre: **DesactualizaciónPorFormato** ***Semántica:*** Chequea si un dato se encuentra desactualizado en base a reglas sintácticas que establecen cuando un dato es actual y cuando no, en base al formato vigente de su tipo de dato. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***NombreSugerido:*** DesactualizaciónPorFormato(Atr,Formato) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. Formato= Expresión regular o reglas que cumple el formato vigente.

Tabla 16.27: Factor Actualidad (continuación)

Actualidad (Continuación)	
Métricas Específicas	***Nombre:*** **DesactualizaciónPorFormato(TelefonoFijo, FormatoPNN[1])** ***Semántica:*** Chequea si un número de telefonía fija se encuentra desactualizado en base al formato vigente del Plan Nacional de Numeración, que establece que a partir de 2010, todos los teléfono pasaron a tener 8 dígitos. ***BasadaEn:*** DesactualizaciónPorFormato ***PropiedadesConfiguración:*** 1. Atr=TelefonoFijo 2. Formato=(Length(TelefonoFijo)==8)

Tabla 16.28: Factor Oportunidad

Oportunidad	
Definición	Captura la demora que existe entre la actualización de un dato y el momento en el que éste se encuentra disponible para ser utilizado.
Otros nombres	Timeliness (ingl.)
Métricas Genéricas	***Nombre:*** **BoolOportunidadAtributoPorFecha** ***Semántica:*** Indica si el valor actualizado de una instancia de atributo está disponible antes de una fecha límite. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***NombreSugerido:*** BoolOportunidadAtributo(Atr) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. T_f=Fecha límite
	Nombre: **BoolOportunidadAtributoPorIntervalo** ***Semántica:*** Indica si el valor actualizado de una instancia de atributo está disponible dentro de un intervalo de vigencia. ***Granularidad:*** instanciaAtributo. ***TipoResultado:*** Boolean. ***NombreSugerido:*** BoolOportunidadAtributoPorIntervalo(Atr) ***PropiedadesConfiguración:*** 1. Atributo=Atr 2. Intervalo=[T_i,T_f]

Tabla 16.29: Factor Oportunidad (continuación)

Oportunidad (continuación)	
Métricas Genéricas	***Nombre:*** **BoolOportunidadEntPorFecha** ***Semántica:*** Indica si una instancia de entidad con sus datos actualizados está disponible antes de una fecha límite. La fecha límite puede ser única o dependiente de cada instancia de la entidad. ***Granularidad:*** instanciaEntidad. ***TipoResultado:*** Boolean. ***NombreSugerido:*** BoolOportunidadEntPorFecha(Ent) ***PropiedadesConfiguración:*** 1. Ent: Entidad 2. Atr: Atributo con la fecha de creación/modificación de la entidad 3. T_f=Fecha(s) límite(s) 4. Referencial: Conjunto de datos extraídos del mundo real o de otra colección de datos que permiten calcular la(s) fecha(s) límite(s).
	Nombre: **BoolOportunidadEntPorIntervalo** ***Semántica:*** Indica si una instancia de entidad está disponible con sus datos actualizados dentro de un intervalo de vigencia. El intervalo de vigencia puede ser único o dependiente de cada instancia de la entidad. ***Granularidad:*** instanciaEntidad. ***TipoResultado:*** Boolean. ***NombreSugerido:*** BoolOportunidadEntPorIntervalo(Ent) ***PropiedadesConfiguración:*** 1. Ent: Entidad 2. Atr: Atributo con la fecha de creación/modificación de la entidad 3. Intervalo(s)=[T_i,T_f] 4. Referencial: Conjunto de datos extraídos del mundo real o de otra colección de datos que permiten calcular el o los intervalos.

El Ejemplo 65 presenta problemas de oportunidad.

Ejemplo 65

En un sistema de emisión de alertas meteorológicas, se envían mensajes hacia todos los suscriptores advirtiendo sobre situaciones climáticas peligrosas, con una hora de inicio y una hora de fin de la alerta. Se constata que muchos mensajes llegan a los suscriptores en un tiempo posterior a la hora de inicio de la alerta.

Referencias

[Abe15] Ziawasch Abedjan, Lukasz Golab y Felix Naumann. «Profiling relational data: a survey». En: *The VLDB Journal—The International Journal on Very Large Data Bases* 24.4 (2015), págs. 557-581.

[AGE18] AGESIC. *Vocabulario de Persona*. Inf. téc. 1.0. Dic. de 2018. URL: https://catalogodatos.gub.uy/dataset/agesic-vocabulario-de-persona (visitado 20-12-2019).

[Ako07] Jacky Akoka, Laure Berti-Equille, Omar Boucelma, Mokrane Bouzeghoub, Isabelle Comyn-Wattiau, Mireille Cosquer, Virginie Goasdoué-Thion, Zoubida Kedad, Sylvaine Nugier, Verónika Peralta y col. «A Framework for Quality Evaluation in Data Integration Systems.» En: *ICEIS (3)*. 2007, págs. 170-175.

[Ber12] Lopez Vazquez Bernabé Poveda. *Fundamento de las Infraestructuras de Datos Espaciales*. UPM Press, 2012.

[Boy11] Isabelle Boydens. «Strategic Issues Relating to Data Quality for E-Government: Learning from an Approach Adopted in Belgium». En: *Practical Studies in E-Government: Best Practices from Around the World*. Ed. por Saïd Assar, Imed Boughzala e Isabelle Boydens. New York, NY: Springer New York, 2011. ISBN: 978-1-4419-7533-1. DOI: 10.1007/978-1-4419-7533-1_7.

[BS16] Carlo Batini y Monica Scannapieco. *Data and Information Quality*. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-24106-7.

[Bus16] IBM Institute for Business Value. *The Chief Data Officer playbook. Creating a game plan to sharpen your digital Edge*. Inf. téc. IBM Institute for Business Value, 2016. URL: https://www.ibm.com/downloads/cas/0OGWOGAW.

[Cab08] Ismael Caballero, Angélica Caro, Coral Calero y Mario Piattini. «IQM3: Information Quality Management Maturity Model». En: *Journal of Universal Computer Science* 14.22 (dic. de 2008), págs. 3658-3685.

[Cha05] Arthur D Chapman. *Principles of data quality*. GBIF, 2005.

[Etc08] Lorena Etcheverry, Verónika Peralta y Mokrane Bouzeghoub. «Qbox-foundation: a metadata platform for quality measurement». En: *proceeding of the 4th Workshop on Data and Knowledge Quality (QDC'2008)*. 2008.

[Fox94] Christopher Fox, Anany Levitin y Thomas Redman. «The notion of data and its quality dimensions». En: *Information Processing & Management* 30.1 (1994), págs. 9-19. ISSN: 0306-4573. DOI: http://dx.doi.org/10.1016/0306-4573(94)90020-5.

[Hea18] Health Information and Quality Authority. *Background paper to support guidance for a data quality framework*. en. Inf. téc. Oct. de 2018, pág. 86.

[IDE18] IDEuy. *Especificación técnica: Calidad de la Información Geográfica*. Inf. téc. Dic. de 2018.

[Int09] DAMA International. *The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK)*. Technics Publications, 2009. ISBN: 978-1935504009.

[Int17] DAMA International. *DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition*. Technics Publications, 2017. ISBN: 978-1634622349.

[ISO08] ISO/IEC. *ISO/IEC 25012:2008 - Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model*. Estándar. International Organization for Standardization (ISO), dic. de 2008. URL: https://www.iso.org/standard/35736.html.

[ISO11] ISO. *ISO/TS 8000-1:2011*. Estándar. International Organization for Standardization (ISO), dic. de 2011. URL: https://www.iso.org/standard/50798.html.

[ISO13] ISO. *ISO 19157:2013 - Geographic information – Data quality*. Estándar. International Organization for Standardization (ISO), dic. de 2013. URL: https://www.iso.org/standard/32575.html.

[ISO15] ISO/IEC. *ISO/IEC 25024:2015 - Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality*. Estándar. International Organization for Standardization (ISO), oct. de 2015. URL: https://www.iso.org/standard/35749.html.

[Lee14] Yang Lee, Stuart E Madnick, Richard Y Wang, Forea Wang y Hongyun Zhang. «A cubic framework for the chief data officer: Succeeding in a world of big data». En: (2014).

[MA19] Melody Chien y Ankush Jain. *Magic Quadrant for Data Quality Tools*. Inf. téc. Gartner, mar. de 2019.

[Nee05] M Pamela Neely. «The product approach to data quality and fitness for use: a framework for analysis». En: *Proc. of 10th International Conference on Information Quality* (2005).

[Ols03] Jack E Olson. *Data quality: the accuracy dimension*. Morgan Kaufmann, 2003.

[OMG11] OMG. *Business Process Model and Notation (BPMN)*. Standard. Ver. 2.0. Object Management Group, 2011.

[Pia18] Mario Piattini, Ismael Caballero, Ana Gómez y Fernando Gualo. *Calidad de Datos*. RA-MA EDITORIAL, 2018. ISBN: 978-84-9964-750-0.

[Pul16] Venkata Sai Venkatesh Pulla, Cihan Varol y Murat Al. «Open Source Data Quality Tools: Revisited». En: *Information Technology: New Generations*. Ed. por Shahram Latifi. Cham: Springer International Publishing, 2016, págs. 893-902. ISBN: 978-3-319-32467-8.

[SC02] Monica Scannapieco y Tiziana Catarci. «Data quality under a computer science perspective». En: *Archivi & Computer* 2 (2002), págs. 1-15.

[Str97] Diane M. Strong, Yang W. Lee y Richard Y. Wang. «Data Quality in Context». En: *Commun. ACM* 40.5 (mayo de 1997), págs. 103-110. ISSN: 0001-0782. DOI: 10.1145/253769.253804.

[TB98] Giri Kumar Tayi y Donald P. Ballou. «Examining Data Quality». En: *Commun. ACM* 41.2 (feb. de 1998), págs. 54-57. ISSN: 0001-0782. DOI: 10.1145/269012.269021.

[Tep17] Jaak Tepandi, Mihkel Lauk, Janar Linros, Priit Raspel, Gunnar Piho, Ingrid Pappel y Dirk Draheim. «The Data Quality Framework for the Estonian Public Sector and Its Evaluation». En: *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXV*. Ed. por Abdelkader Hameurlain, Josef Küng, Roland Wagner, Sherif Sakr, Imran Razzak y Alshammari Riyad. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017, págs. 1-26. ISBN: 978-3-662-56121-8. DOI: 10.1007/978-3-662-56121-8_1.

[Wen07] Kristin Wende. «A model for data governance-Organising accountabilities for data quality management». En: *ACIS 2007 Proceedings* (2007), pág. 80.

[WS96] Richard Y. Wang y Diane M. Strong. «Beyond Accuracy: What Data Quality Means to Data Consumers». En: *J. Manage. Inf. Syst.* 12.4 (mar. de 1996), págs. 5-33. ISSN: 0742-1222.